

# **Comparative Analysis of Connection and Disconnection in the Human Brain Using Diffusion MRI: New Methods and Applications**

*Jonathan D. Clayden*



Doctor of Philosophy  
Institute for Adaptive and Neural Computation  
School of Informatics  
University of Edinburgh  
2008



# Abstract

**D**IFFUSION magnetic resonance imaging (dMRI) is a technique that can be used to examine the diffusion characteristics of water in the living brain. A recently developed application of this technique is “tractography”, in which information from brain images obtained using dMRI is used to reconstruct the pathways which connect regions of the brain together. Proxy measures for the integrity, or coherence, of these pathways have also been defined using dMRI-derived information.

The “disconnection hypothesis” suggests that specific neurological impairments can arise from damage to these pathways as a consequence of the resulting interruption of information flow between relevant areas of cortex. The development of dMRI and tractography have generated a considerable amount of renewed interest in the disconnectionist thesis, since they promise a means for testing the hypothesis *in vivo* in any number of pathological scenarios. However, in order to investigate the effects of pathology on particular pathways, it is necessary to be able to reliably locate them in three-dimensional dMRI images.

The aim of the work described in this thesis is to improve upon the robustness of existing methods for segmenting specific white matter “tracts” from image data, using tractography, and to demonstrate the utility of the novel methods for the comparative analysis of white matter integrity in groups of subjects.

The thesis begins with an overview of probability theory, which will be a recurring theme throughout what follows, and its application to machine learning. After reviewing the principles of magnetic resonance in general, and dMRI and tractography in particular, we then describe existing methods for segmenting particular tracts from group data, and introduce a novel approach. Our innovation is to use a reference tract to define the topological characteristics of the tract of interest, and then search a group of “candidate” tracts in the target brain volume for the best match to this reference. In order to assess how well two tracts match we define a heuristic but quantitative tract similarity measure.

In later chapters we demonstrate that this method is capable of successfully segmenting tracts of interest in both young and old, healthy and unhealthy brains; and then describe a formalised version of the approach which uses machine learning methods to match tracts from different subjects. In this case the similarity between tracts is represented as a matching probability under an explicit model of topological variability between equivalent tracts in different brains. Finally, we examine the possibility of comparing the integrity of groups of white matter structures at a level more fine-grained than a whole tract.





## Acknowledgements

I HAVE been extremely fortunate to be supervised during my PhD by Dr Mark Bastin and Dr Amos Storkey in Edinburgh, UK; and by Prof. David Laidlaw at Brown University in Providence, RI, USA. Their fervour has been contagious; their guidance and support have been invaluable; and, fortunately for me, Mark's aim with a screwed-up paper ball has not improved noticeably over the past three years. I am also grateful to David for being so welcoming during my stay in the United States.

The credit—or perhaps, the blame—for allowing this computer scientist into the domain of clinical neuroscience lies squarely with the Neuroinformatics Doctoral Training Centre (DTC) here in Edinburgh. It is no exaggeration to say that without the DTC this thesis would not exist. Not only has it supplied my funding—ultimately from the EPSRC and MRC research councils—but it has also served as a forum for discussing and hearing ideas from across the diverse scope of neuroinformatics. My thanks therefore go to all the academics, students and administration staff for making the DTC what it is; particularly to the directors, Prof. David Willshaw, Dr Mark van Rossum and Dr Jim Bednar, and to Pat Ferguson, who has been extremely accommodating and helped to keep me properly organised.

The Brain Imaging Centre at Edinburgh's Western General Hospital has been a second home—although, fortunately, only metaphorically speaking—over the last three years; and once again, the atmosphere has been friendly and constructive. My gratitude is due to the radiographers in the centre for helping with scanning, and to the other PhD students in medical physics for regular opportunities to discuss all things random and eat cake.

A special mention must go to Dr Paul Armitage, for critically reading papers and this thesis, and for ribbing me far beyond the call of duty. Many thanks also to the other souls who have worked their way through part or all of what follows, and made constructive comments: namely, Mark Bastin, Ann Clayden, Susana Muñoz Maniega, Amos Storkey and Terry Thorpe. And, of course, without the input of my examiners, Dr Daniel Alexander and Prof. Ian Marshall, all this effort would have been to little purpose—so I thank them for their time and helpful feedback.

My parents have both been very supportive, and have tried to take an interest in what I've been doing, despite both being humanities graduates! I would like to extend my thanks to them, therefore, and to the rest of my family and friends. Finally, I reserve special affection and admiration for Susana, *musa mía*.

JDC

Edinburgh, 27 Sept. 2007

London, 31 Jan. 2008



## Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

*(Jonathan D. Clayden)*



Dedicated to the memory of Gordon Dew, “Grandad” (1920–1997).



The ecstatic love of a young writer for the old writer he will be some day is ambition in its most laudable form. This love is not reciprocated by the older man in his larger library, for even if he does recall with regret a naked palate and a rheumless eye, he has nothing but an impatient shrug for the bungling apprentice of his youth.

—Vladimir Nabokov





---

# Table of contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Problem statement . . . . .	6
<b>2</b>	<b>Probability and machine learning principles</b>	<b>9</b>
2.1	Fundamentals of probability theory . . . . .	9
2.2	Probability distributions . . . . .	10
2.3	Inference and learning . . . . .	12
2.4	Maximum likelihood . . . . .	14
2.5	Expectation–Maximisation . . . . .	16
2.6	Sampling methods . . . . .	19
2.6.1	Rejection sampling . . . . .	20
2.6.2	Markov chain Monte Carlo . . . . .	21
2.7	Summary . . . . .	23
<b>3</b>	<b>Magnetisation, excitation and relaxation</b>	<b>25</b>
3.1	State and spin . . . . .	25
3.2	Protons in a magnetic field . . . . .	28
3.3	The NMR signal . . . . .	29
3.4	Pulse sequences . . . . .	31
3.5	On ghosts and pile-ups . . . . .	34
3.6	Summary . . . . .	35
<b>4</b>	<b>Diffusion magnetic resonance imaging</b>	<b>37</b>
4.1	The Einstein picture . . . . .	37
4.2	Diffusion tensor imaging . . . . .	39
4.3	A more general displacement distribution . . . . .	41
4.4	The role of registration . . . . .	42
4.5	Diffusion MRI in the clinic . . . . .	44
4.6	Summary . . . . .	45
<b>5</b>	<b>White matter fibre tracking</b>	<b>47</b>
5.1	Streamlines . . . . .	47
5.2	BEDPOST . . . . .	51
5.3	Fast marching . . . . .	53
5.4	High angular resolution methods . . . . .	54
5.4.1	Using $\mathbf{q}$ -space . . . . .	55
5.4.2	Spherical deconvolution . . . . .	57
5.5	Applications and challenges . . . . .	58
5.6	Summary . . . . .	60

<b>6</b>	<b>Neighbourhood tractography</b>	<b>61</b>
6.1	Group comparison in white matter . . . . .	61
6.2	Tract-specific comparison . . . . .	63
6.3	Similarity and matching . . . . .	65
6.3.1	The reduced tract . . . . .	67
6.3.2	A similarity measure . . . . .	67
6.4	Validation and application . . . . .	68
6.5	How many seeds? . . . . .	72
6.6	Evaluation of the similarity measure . . . . .	75
6.7	The next step . . . . .	76
<b>7</b>	<b>Applications</b>	<b>77</b>
7.1	Tractography in the ageing brain . . . . .	77
7.2	Old versus young . . . . .	81
7.3	Improving the reference tracts . . . . .	82
7.4	A schizophrenia study . . . . .	83
7.5	Conclusions . . . . .	85
<b>8</b>	<b>Model-based tract matching</b>	<b>87</b>
8.1	B-splines . . . . .	87
8.2	Tract representation revisited . . . . .	89
8.3	Comparing spline tracts . . . . .	91
8.4	Training and using the model . . . . .	93
8.5	Advantages and limitations . . . . .	96
8.6	An unsupervised approach . . . . .	98
8.7	Conclusions . . . . .	102
<b>9</b>	<b>Anisotropy profiling</b>	<b>103</b>
9.1	A single profile . . . . .	103
9.2	The median tube . . . . .	105
9.3	Intersubject tube alignment . . . . .	106
9.4	Comparative profiling . . . . .	107
9.5	Discussion . . . . .	109
<b>10</b>	<b>Conclusions</b>	<b>111</b>
10.1	Tract segmentation . . . . .	111
10.2	Comparative analysis . . . . .	113
10.3	Final remarks . . . . .	113
<b>A</b>	<b>List of abbreviations</b>	<b>115</b>
<b>B</b>	<b>Summary of publications</b>	<b>117</b>
	<b>Bibliography</b>	<b>119</b>

---

## A note on notation

THIS THESIS draws on elements of theory from clinical and nonclinical neuroscience, statistics, physics and computer science; and as such it has been extremely difficult to maintain consistency of notation throughout. Each field has its own conventions, and they are often incompatible. It has therefore been necessary to reuse some notation—particularly adornments such as the hat (^), tilde (~) and asterisk (\*)—and occasionally to depart from standard nomenclature. Short range (i.e. within-chapter) consistency has generally been favoured over long range consistency where a choice needed to be made.

Nevertheless, we have endeavoured to stick to certain basic principles of notation. Scalar variables and sets are generally labelled with italic Latin or Greek letters, as in  $x$  or  $\phi$ . A vector or tuple is written using bold notation, as in  $\mathbf{r}$ , except in the sections on quantum mechanics in chapter 3, which use the ket notation,  $|r\rangle$ . Matrices—and vectors, when they are used as single column or single row matrices—are written using sans-serif font ( $\mathbf{M}$ ). In chapter 2, we use calligraphic notation, as in  $\mathcal{A}$ , to represent the sample space of a random variable.

In addition we sometimes use the shorthand  $\{x_i\}$  to mean the set of values of  $x_i$  for all appropriate values of  $i$ ; or  $\{1..N\}$  to mean the set of integers between (and including) 1 and  $N$ . Vector literals are written using bracket notation, as in  $\mathbf{x} = (1, 2, 3)$ ; and  $(x_i)$  is the shorthand for the indexed components of a vector. Matrix literals are written using square bracket notation, like

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

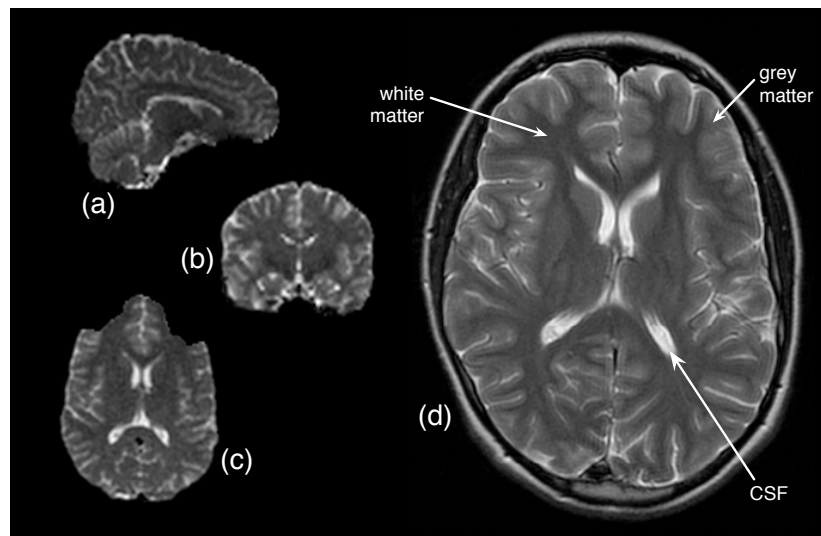
Familiarity with common notation for operations on sets, vectors and sequences will be assumed, as will a grasp of basic calculus.

Magnetic resonance images are generally shown as two-dimensional slices, using the radiological convention whereby the subject's left side appears on the right of the image.



Almost simultaneously, in the mid-1960s, the foundations of a method called diffusion magnetic resonance imaging (dMRI) were being laid, a technique which could be used to characterise the diffusion of water in living tissue. This technique, in common with all magnetic resonance imaging (MRI) methods, made use of an earlier discovery about the behaviour of certain types of particles in a very strong magnetic field: the nuclear magnetic resonance

[illegible]



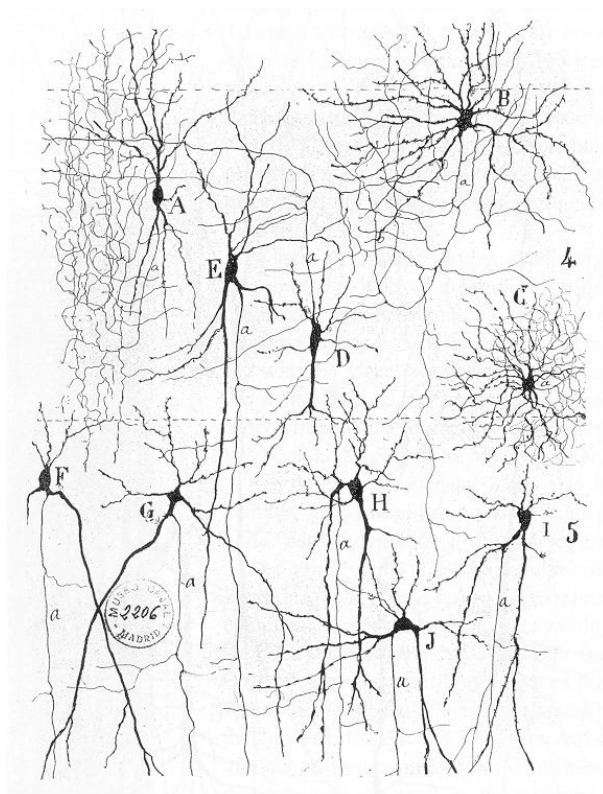
**Figure 1.2:** Magnetic resonance images of the brain, shown in sagittal (a), coronal (b) and axial (c) planes, perpendicular to the left–right, anterior–posterior and superior–inferior directions respectively. The high resolution axial image (d) shows clear contrast between the three main tissue types. By radiological convention, the left side of the brain from the patient's perspective is shown on the right side of all images.

(NMR) phenomenon, which had been used for chemical analysis for some time before imaging methods reached maturity. Both NMR and MRI won their respective pioneers Nobel prizes. Felix Bloch and Edward Mills Purcell shared the 1952 physics prize for their work on the former; while Paul Lauterbur and Peter Mansfield were awarded the prize in medicine in 2003 for developing MRI, despite outstanding controversy over whether they were truly the first to demonstrate the technique.

MRI is now routinely used to create images of almost every part of the body for clinical diagnosis and prognosis, but it is particularly valuable for imaging the brain, whose details are obscured for x-rays by the bone of the skull. (The major regions, or *lobes*, of the brain are shown relative to the skull in Fig. 1.1.) Since the method involves no ionising radiation, it can also be used repeatedly on a single subject without fear of tissue damage. Imaging using magnetic resonance not only allows clinicians and researchers to visualise brain structure at a respectable resolution—on the order of 1 mm in each dimension—it can also be tailored to enhance contrasts between different tissue types, or between healthy and unhealthy tissue. In Fig. 1.2, for example, the distinction is quite clear between grey matter, white matter, and the cerebrospinal fluid (CSF) in which the brain is bathed. With diffusion MRI, image contrast is related to the local magnitude of water diffusion.

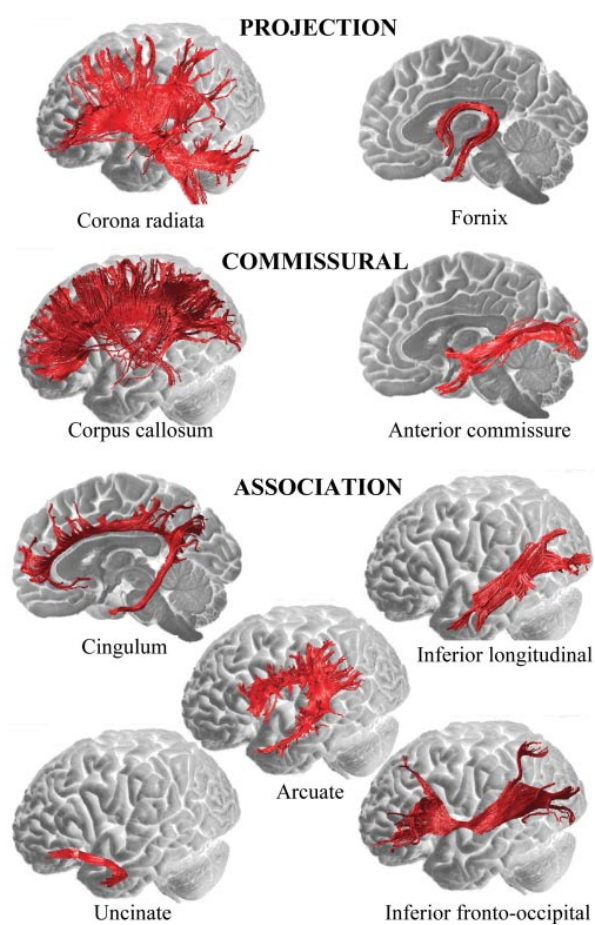
The potential of dMRI for studying white matter in particular was not immediately realised, and it was not until Peter Basser, James Mattiello and Denis Le Bihan described a way of measuring not just the magnitude but also the orientational structure of diffusion using NMR (Basser *et al.*, 1994a)—a method called diffusion tensor imaging—that this potential began to be fully realised. With this new development established, methods for virtual reconstruction of white matter structures—or **tractography**—quickly followed.

It has been possible for decades to examine the structure of connective tissue at the individual neuron level. Santiago Ramón y Cajal, often called one of the fathers of modern neuroscience, created superbly detailed and quite beautiful drawings of a great variety of complex neurons more than a hundred years ago (see Fig. 1.3), thereby earning him, and the inventor of the staining technique he used, Camillo Golgi, the 1906 Nobel prize for medicine. Moreover, histological methods have since improved to the point where tracing the routes of axons—the projections of neurons which bundle into larger connective structures—can be performed effectively and with impressive accuracy. Nonetheless, the development of a method for probing the connectivity of *living* brains was no small achievement, despite its far coarser



**Figure 1.3:** Drawing of Golgi-stained neurons in auditory cortex by Santiago Ramón y Cajal. A considerable variety of cell morphologies is visible. Reproduced from *Texture of the nervous system of man and the vertebrates*, translated and edited by Pedro and Tauba Pasik.

**Figure 1.4:** Illustration of some major white matter fasciculi, visualised using tractography and categorised after Meynert. Reproduced from Catani & ffytche (2005).



resolution, since all the alternatives oblige the researcher to wait for his subject to die, or to study animals which can be sacrificed. For the first time, it may be possible to test the disconnection hypothesis in patients with appropriate disorders.

Theodor Meynert, a nineteenth century neuropathologist, was the first to distinguish white matter structures, or **fasciculi**, into projection fibres, which connect cortical and subcortical grey matter together; commissural fibres, which link the two brain hemispheres; and association fibres, which connect distal cortical regions within a hemisphere. Fig. 1.4 shows examples of important tracts in each category. The corpus callosum, in the commissural class, is the largest white matter structure in the brain, connecting all the main lobes between hemispheres. Whatever principle one uses to categorise the various fasciculi, it is to be expected that interrupting different types of connection will have different effects; and so, conversely, it might be anticipated that different diseases affect different fasciculi. Indeed, Meynert described psychiatry as simply the study of diseases of the forebrain.

Despite its relative immaturity, tractography already offers the possibility of examining tract-specific effects of disease during the course of the illness; and at some point in the future it may become possible to use this kind of dMRI-derived information to inform the prognosis of patients with white matter diseases. At present, however, the robust location and characterisation of specific white matter tracts between subjects remains elusive. Tractography algorithms are able to segment particular tracts, but they are typically very strongly dependent on their initialisation, and principles for guiding the choice of starting condition are lacking. This thesis aims to take steps in that direction.

## 1.1 Problem statement

In order to establish whether a certain disease may be detrimentally affecting a particular white matter structure, it is typically constructive to compare the MR-visible characteristics of healthy and unhealthy examples of the tract in question. Since there is usually substantial variability in such characteristics even between normal individuals, due to imaging noise and genuine biological disparity, the comparison needs to be performed statistically between groups of subjects with similar clinical statuses. If a clear distinction is found, on aggregate, then this information can be used to estimate how likely it is that a new subject falls into one or other of the groups.

Under these circumstances, it is important to strive to minimise the impact of uncontrolled factors which may mask or exaggerate the true differences between the groups (or lack thereof). In statistical terms, an additional source of variance within the groups may lead to a falsely negative outcome, while a similar effect between groups may produce a falsely positive conclusion, suggesting that there is a group difference when in fact there is not.

One potentially large source of variance in the comparative analysis of white matter comes from segmentation. In order to compare a particular tract between groups, one must first identify it in each individual brain volume; and this should be done as consistently as possible to avoid introducing bias. Secondly, once comparable tracts have been identified in each brain volume, it is desirable that measures used to quantify differences between the groups be sensitive to white matter degradation whilst being relatively invariant to other nuisance factors. Improvement of segmentation consistency and examination of within-group and between-group variability in dMRI-derived measures of white matter integrity are the joint aims of the new work described in chapters 6–9.

The structure of the thesis is as follows. After outlining the general principles of probability (in chapter 2) and the physics of nuclear magnetic resonance (in chapter 3), we go on to discuss the nature of water diffusion in the brain. It is also explained, in chapter 4, how and why dMRI can be useful for probing white matter structure. Chapter 5 provides a survey of the tractography literature, thereby giving a sense of the potential diversity of segmentation methods, even within the general fibre tracking approach. Tractography is compared to other segmentation approaches in the early part of chapter 6, after which we begin to describe our novel take on the problem, whereby a reference tract is defined in advance to epitomise the topology of the white matter structure of interest, and a segmentation is chosen from among a



number of candidates in each brain volume by comparing them algorithmically to the reference and selecting the best match. The technique is demonstrated on healthy young and (in chapter 7) aged and unhealthy subjects and shown to improve segmentation consistency compared to a simpler alternative strategy. Certain limitations come to light, however, and so in chapter 8 the method is further developed, and its principles are formalised using a probabilistic model. Techniques from machine learning are then applied to fitting the model parameters from data and performing segmentation in complete data sets. Finally, in chapter 9, we describe an attempt to compare a proxy measure for tract integrity between groups at a fine spatial scale, and investigate how the measure varies between and within populations.



## 2

# Probability and machine learning principles

PROBABILITY will be a recurring theme throughout the thesis. Indeed, its influence is surprisingly pervasive in the material making up the remaining chapters. The quantum mechanical underpinnings of the NMR effect which are outlined in chapter 3 describe probabilistic behaviours; the diffusion of water in the brain measured by dMRI is fundamentally a stochastic process; probabilistic sampling techniques are important to some of the tractography methods described in chapter 5; and the machine learning methods that we apply to the problem of tract selection in chapter 8 are probabilistic by their nature.

In this chapter we lay out the theory of probability and describe those machine learning and inference methods upon which later chapters are dependent. General references for this material include MacKay (2003) and Bishop (2006).

## 2.1 Fundamentals of probability theory

Consider a nondeterministic experiment, such as rolling a fair die. The result of this experiment on any given trial will be one of exactly six possibilities, representing the number of spots on the uppermost face of the die. Moreover, each of these possibilities is equally likely; so over a very large number of trials, all six will occur an approximately equal number of times. This kind of experiment is represented mathematically by a random variable, which we call  $X$ . The set of possible outcomes, or *sample space*, relating to  $X$  is  $\{1, 2, 3, 4, 5, 6\}$ . The probability of each of these outcomes on a single trial is, of course,  $1/6$ .

In general, we denote the sample space for a discrete random variable,  $X$ , as  $\mathcal{A}_X = \{a_i\}$ , where each member of the set has a corresponding probability,  $p_i$ . We write

$$\Pr(x = a_i) = p_i ,$$

where “Pr” represents “the probability that”, and  $x$  represents a particular outcome. The result  $x = a_i$  is an example of an *event*, a concept which can generally encapsulate the occurrence of any subset of the sample space: if  $E$  is some subset of  $\mathcal{A}_X$ , we have

$$\Pr(E) = \Pr(x \in E) = \sum_{a_i \in E} \Pr(x = a_i) . \quad (2.1)$$

In the example of the die, if  $E = \{1, 2\}$ , then  $\Pr(E)$ —the probability that the outcome of a trial is *either* 1 or 2—is the sum of  $\Pr(x = 1)$  and  $\Pr(x = 2)$ , i.e.  $1/3$ .

The basic axioms of probability state that the probability of any event is greater than or equal to zero, with the latter representing an impossible event; and that the probability of the whole sample space is unity—i.e. every outcome must be drawn from the space. That is,

$$\forall E \subseteq \mathcal{A}_X . \Pr(E) \geq 0 \quad \Pr(\mathcal{A}_X) = 1 . \quad (2.2)$$

Naturally, Eq. (2.2) additionally implies that  $\Pr(E) \leq 1$ . In general, given any pair of events,  $E_1$  and  $E_2$ , the probability of their union is given by

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2) . \quad (2.3)$$

This third axiom follows straight from Eq. (2.1). In the special case in which  $\Pr(E_1 \cap E_2) = 0$ , the two events cannot occur simultaneously and are therefore mutually exclusive.

We now consider another experiment, represented by the random variable  $Y$ , which consists of flipping a coin. The sample space for this variable can be represented as  $\mathcal{A}_Y = \{0, 1\}$ , where 0 represents a tail and 1 a head. If we perform both experiments together, what is the probability that the die roll produces a 6 *and* the coin toss gives a head? We represent this **joint probability** as  $\Pr(x = 6, y = 1)$ . Since the roll of the die and the coin toss can be assumed to have no influence on each other, the two events are *independent* and the joint probability is simply the product of the individual probabilities. For the case of two events that are not independent, we need to introduce the concept of a **conditional probability**, which is defined by

$$\Pr(x = a_i | y = b_j) \equiv \frac{\Pr(x = a_i, y = b_j)}{\Pr(y = b_j)} \quad \text{if } \Pr(y = b_j) \neq 0 ,$$

and should be interpreted as “the probability that  $x = a_i$  given that  $y = b_j$ ”. Hence, if we omit the particular value of each outcome to indicate the general case, it follows by trivial rearrangement that

$$\Pr(x, y) = \Pr(x | y) \Pr(y) , \quad (2.4)$$

which is called the *product rule* for probabilities. Consequently, the following are equivalent statements of independence between  $X$  and  $Y$ :

$$\Pr(x, y) = \Pr(x) \Pr(y) \quad \Pr(x | y) = \Pr(x) .$$

Finally, given a group of joint probabilities,  $\Pr(x, y)$ , we can calculate the so-called **marginal probability**,  $\Pr(x)$ , by summing over all possible values of  $y$ ; an operation known as marginalisation:

$$\Pr(x) \equiv \sum_{y \in \mathcal{A}_Y} \Pr(x, y) .$$

It follows from Eq. (2.4) that

$$\Pr(x) = \sum_{y \in \mathcal{A}_Y} \Pr(x | y) \Pr(y) , \quad (2.5)$$

a relationship which is called the *sum rule* for probabilities. These basic rules for combining probabilities together are extremely important in machine learning.

## 2.2 Probability distributions

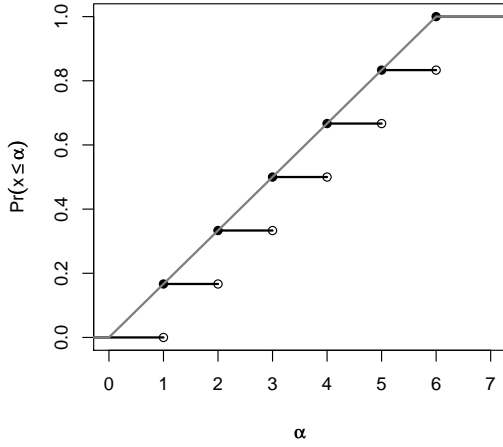
Every random variable has associated with it a probability distribution, which can be used to assign to some interval  $[\alpha, \beta]$  over the set of real numbers,  $\mathbb{R}$ , a probability that the corresponding outcome will fall within that interval on any given trial. For the example of our fair die, the distribution is easily defined as

$$P(x) = \begin{cases} \Pr(x) & \text{if } x \in \mathcal{A}_X \\ 0 & \text{otherwise.} \end{cases} \quad (2.6)$$

In this case,  $P(x)$  is called the **probability mass function** (p.m.f.) for  $X$ . It then follows from Eqs (2.1) and (2.6) that

$$\Pr(\alpha \leq x \leq \beta) = \sum_{a_i \leq \beta} P(a_i) - \sum_{a_i < \alpha} P(a_i) = \sum_{a_i \leq \beta} \Pr(a_i) - \sum_{a_i < \alpha} \Pr(a_i) ,$$

where  $\Pr(a_i)$  is shorthand for  $\Pr(x = a_i)$ , and  $a_i \in \mathcal{A}_X$  in all cases.



**Figure 2.1:** Cumulative distribution functions for discrete (black) and continuous (grey) uniform distributions. Open and closed circles indicate that each interval with a particular cumulative probability is open at one end and closed at the other; i.e.  $\Pr(x \leq \alpha) = 1/6$  for  $\alpha \in [1, 2)$ , for the discrete case.

It may appear that the distribution function buys us nothing over the individual probabilities for each outcome—after all, its only addition is to make explicit the fact that the probability of an outcome outside of the sample space is zero, a fact which follows uncontroversially from Eq. (2.2). However, the significance of probability distributions is far more obvious when we deal with continuous random variables.

Consider a continuous analogue of the die-rolling scenario, in which the outcome can be *any* real number in the interval  $[0, 6]$ . The distribution for this continuous random variable is now defined by a **probability density function** (p.d.f.); specifically

$$P(x) = \begin{cases} \frac{1}{6} & \text{if } 0 \leq x \leq 6 \\ 0 & \text{otherwise.} \end{cases} \quad (2.7)$$

Notice that this distribution, while similar to the p.m.f. for the discrete case, is nonzero at an infinite number of points. As a result, the value of this p.d.f. at any given point does *not* represent a probability—if it did, the sum of probabilities across the sample space would be infinite, which defies the axioms of Eq. (2.2). Instead, the p.d.f. represents probability *density*, which is related to probability through integration:

$$\Pr(\alpha \leq x \leq \beta) = \int_{\alpha}^{\beta} P(x) dx .$$

Consequently, the probability of an outcome having any particular value—i.e.  $\Pr(x = \alpha)$ —is *zero* for all values of  $\alpha \in \mathbb{R}$ , both within and outside the sample space, for any continuous random variable.

Every interval over the real numbers is a subset of  $\mathbb{R}$ , so the normalisation axiom in Eq. (2.2) implies that

$$\int_{-\infty}^{\infty} P(x) dx = \int_{\mathcal{A}_X} P(x) dx = 1 , \quad (2.8)$$

since that part of the integral that is outside the sample space will be equal to zero.

The difference between the discrete and continuous versions of the distribution are most easily illustrated by comparing their *cumulative distribution functions* (c.d.f.s), which map each real number,  $\alpha$ , to the probability that  $x$  is less than or equal to  $\alpha$ . These functions are shown graphically in Fig. 2.1. It can be seen that  $\Pr(x \leq \alpha)$  increases in jumps for the discrete case and smoothly for the continuous case; but for all integer values of  $\alpha$ , the value of the c.d.f. is the same in both cases. Note that the c.d.f. is zero for all values below the lower bound of the sample space, and unity for all values above its upper bound, in each case.

Probability distributions are not only used in relation to static events—it is also common to consider a sequence of random variables,  $(X(t))$ , which are parameterised by  $t$ , often representing time in some sense. This parameter may be discrete or continuous. Such a collection of related variables, used to represent the state of some time-dependent system, is called a *stochastic process*. The evolution of such a process over time is then described by conditional distributions, such as  $P(X(t)|X(t-1), X(t-2), \dots)$  for the discrete-time case.

A final foundational concept with regard to probability distributions is that of the **expectation** of a random variable, which is essentially a weighted mean value over the sample space. For a discrete random variable,  $X$ , it is defined as

$$\langle X \rangle = \sum_{x \in \mathcal{A}_X} x P(x), \quad (2.9)$$

and equivalently, using an integral, for the continuous case. Note that the expectation is a property of the random variable—or equivalently, its distribution—rather than of any outcome. We can also find the expectation of a function of  $X$  with respect to its probability distribution:

$$\langle f(X) \rangle = \sum_{x \in \mathcal{A}_X} f(x) P(x). \quad (2.10)$$

If we know the distribution of a particular random variable, we can deduce the distribution of other random variables related to it. Let us assume that  $X \sim U(0, 6)$ , which is shorthand to say that  $X$  is *uniformly* distributed over the sample space  $[0, 6]$ , as described by Eq. (2.7). We now wish to know the distribution of the random variable

$$Y = 1 - \sqrt{X}.$$

We cannot find the distribution of  $Y$  by simply mapping the sample space accordingly, because this nonlinear function of  $X$  cannot be expected to have a uniform distribution itself; and even if it were linear, we would still need to ensure that the new distribution remains properly normalised. Instead, the rules of integration by substitution (Riley *et al.*, 2002) tell us that, using the Leibniz notation,

$$dx = \frac{\partial x}{\partial y} dy = 2(y-1) dy;$$

so from Eq. (2.8),

$$\int_0^6 \frac{1}{6} dx = \int_1^{1-\sqrt{6}} \frac{2(y-1)}{6} dy = \int_{1-\sqrt{6}}^1 \frac{1-y}{3} dy = 1.$$

The distribution for  $Y$  is thus  $P(y) = (1-y)/3$ , and the sample space is  $\mathcal{A}_Y = [1 - \sqrt{6}, 1]$ . It should be noted that substitutions for functions of more than one original variable are more complex, requiring the calculation of a Jacobian matrix of partial derivatives.

This process of finding the distribution of one random variable from that of another is very important when artificially sampling from a distribution. We sometimes wish to generate data with a certain distribution without truly sampling the value of an appropriate random variable many times; and while computing environments typically provide a method to generate uniformly distributed pseudorandom numbers, an appropriate transformation is needed to turn these into samples from the distribution of interest.

## 2.3 Inference and learning

So far we have talked about probabilities in terms of the chance of a particular event happening, on average, as a result of running a trial of a particular experiment. This interpretation of probability is the classical *frequentist* interpretation. However, there is an alternative, and broader, interpretation of probability which includes the sense of a *degree of belief*. Consider, for example, the relationship between the fact that the sky is cloudy and the fact that it is raining. Intuitively, if we are told that the sky is cloudy then it seems much more likely

that it is raining than if we are told that the sky is clear, or if we know nothing at all about state of the sky. However, the proposition “it is raining” cannot be strictly represented by a random variable since the experiment required to find an outcome (for example, going outside to look) is deterministic. Either it is raining or it isn’t—there can be no two ways about it. It is also unrepeatable, since it is fixed to a particular time and we cannot sample the state of the weather *right now* many times. However, if we allow the broader interpretation of probability, we can admit a conditioned probability  $\Pr(\text{raining}|\text{cloudy})$ , which represents how strongly we believe our proposition, given the truth of another proposition which says “the sky is cloudy”. Moreover, we can use a distribution over the state space, in this case {raining, not raining}, to encapsulate the uncertainty we have about the proposition.

If this talk of using some propositions to inform others sounds like logical deduction, it is no coincidence. Some authors who subscribe to this broader, *Bayesian*, interpretation of probability—notably Jaynes (2003)—have been keen to frame it as a form of logical framework for the uncertain propositions that are common in science.

Note that before we are told about the state of the sky, it cannot influence our belief of whether it is raining or not. As a result, the **prior probability** that it is raining,  $\Pr(\text{raining})$ , may be assumed to take the value 0.5, indicating total uncertainty. The distribution is then uniform over the two outcomes, which is an *uninformative* prior distribution because it tells us nothing except the size of the state space, which we already know. On the other hand, it may be that assumptions and information unrelated to the sky conditions could be incorporated into the prior distribution. Say, for example, that weather records tell us that it typically rains 20 per cent of the time—in that case we might instead use the prior  $\Pr(\text{raining}) = 0.2$ . This is a trivial case of *inference*, whereby we use sample data—the weather records—to infer the nature of the distribution that is used to predict future weather. Note that we need to make an assumption, that previous weather will be representative of the future, in order to do even this simple an inference. In general, the making of assumptions is a prerequisite for inference.

Let’s say that we have encoded our prior knowledge in a distribution of some kind. Now, introducing the knowledge that it is cloudy will alter the plausibility of the proposition that it is raining, but how? Given the fact that joint probabilities are symmetric, i.e.  $P(x, y) = P(y, x)$ , the relationship between the prior probability and the conditioned **posterior probability** can be established straight from Eq. (2.4). It is

$$\Pr(\text{raining}|\text{cloudy}) = \frac{\Pr(\text{cloudy}|\text{raining})\Pr(\text{raining})}{\Pr(\text{cloudy})}.$$

This relationship is the extremely important result known as **Bayes’ rule**, after the 18th century mathematician and clergyman, the Rev. Thomas Bayes. It is significant because it describes a mathematical way to use relevant information to update the level of belief in a proposition—that is, to *learn*.

It turns out that the rules for manipulating probabilities that we looked at earlier can be applied to probability densities as well as probabilities, although showing that this is the case requires a more formal exploration of probability in terms of measure theory, which is beyond our scope here (see Kingman & Taylor, 1966). The same applies to Bayes’ rule, so we can write in general,

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}. \quad (2.11)$$

The denominator of this equation, known in this context as the **evidence**, is commonly expanded using Eq. (2.5), in which the sum is replaced by an integral for the continuous case:

$$P(y) = \int_{\mathcal{A}_x} P(y|x)P(x). \quad (2.12)$$

At this point, having introduced the Bayesian interpretation of probability, we will drop the notational distinction between distribution variables (including random variables) and outcome variables which has been used so far. This is common practice in the literature, and it helps to reduce the quantity of notation needed for dealing with more complex problems.

## 2.4 Maximum likelihood

We now have the tools in place to consider a more practically interesting example. Let us say that we have a random variable,  $x$ . We suspect that  $x$  is approximately normally distributed; that is,  $x \sim N(\mu, \sigma^2)$ , where  $\mu$  (the mean) and  $\sigma^2$  (the variance) are parameters of the distribution. We do not know what these parameters are, but if we want to make predictions about  $x$  we will need to know them. The definition of the normal, or Gaussian, distribution tells us that

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (2.13)$$

In order to make any progress towards establishing  $\mu$  and  $\sigma$ , we need some information. Let us assume that we have a data set,  $D = \{d_i\}$  for  $i \in \{1..N\}$ , of example values of  $x$ . Since we are working on the assumption that  $x$  has the distribution given above, these data are assumed to be *samples* from the distribution. We assume that each sample has no dependence on any other, and that the values of  $\mu$  and  $\sigma$  did not vary across the sample, a combination called the assumption of *independent and identically distributed* (i.i.d.) data. Hence, the product rule gives us a joint distribution for the whole sample data set:

$$P(D|\mu, \sigma) = \prod_{i=1}^N P(d_i|\mu, \sigma). \quad (2.14)$$

The distribution given in Eq. (2.14) may not appear to get us any closer to an actual estimate for the parameters. But note that, from Eqs (2.11) and (2.12),

$$P(\mu, \sigma|D) = \frac{P(D|\mu, \sigma)P(\mu, \sigma)}{\iint P(D|\mu, \sigma)P(\mu, \sigma) d\mu d\sigma}. \quad (2.15)$$

Note that the distribution  $P(D|\mu, \sigma)$ , which is known as the **likelihood** of the parameters, is meaningful in a frequentist sense, since the elements of the data set are sample outcomes of the random variable  $x$ . However, the prior and posterior distributions over the parameters possess only Bayesian significance, since their values are fixed but unknown.

It makes intuitive sense to use as an estimate of the parameters those values which sit at the mode—that is, the point of maximal probability density—of the posterior distribution  $P(\mu, \sigma|D)$ . This approach amounts to finding the *most likely* values of the parameters in light of the sample data available. If we have no prior information about the parameters, so that  $P(\mu, \sigma)$  is uninformative, then maximising the posterior is equivalent to maximising the likelihood, since the evidence is a normalisation factor that is not dependent on the values chosen for  $\mu$  and  $\sigma$ . Hence, we can find a *maximum likelihood estimator* for the parameters by maximising the value of Eq. (2.14) with respect to them.

In practice, it is often mathematically easier to maximise the (natural) logarithm of the likelihood. This is valid because  $\ln n$  will always increase when  $n$  increases—we say that the logarithm is a *monotonically increasing* function. Elementary calculus tells us that at the maximum of a function its derivative is zero, so from Eqs (2.13) and (2.14), our estimator of  $\mu$  is given when

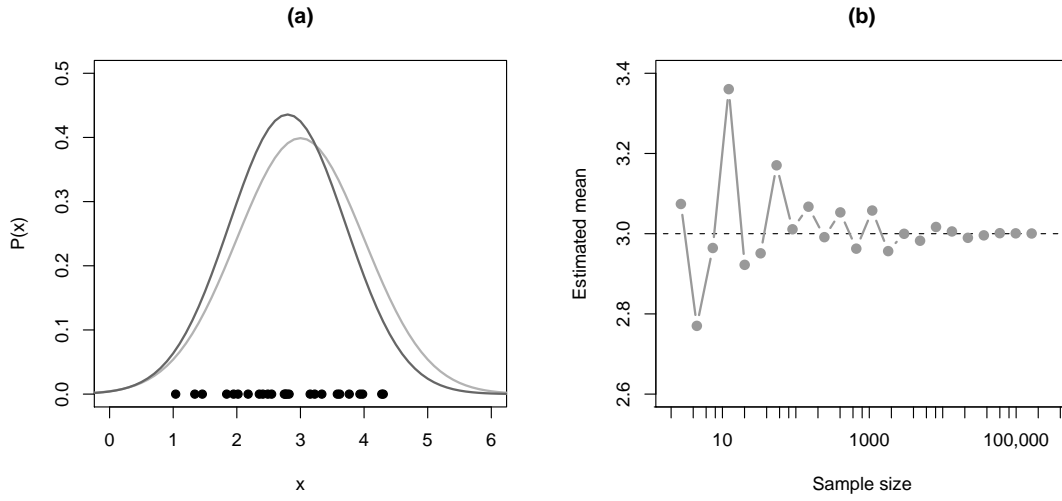
$$\frac{\partial}{\partial \mu} \left( -\frac{1}{2} \sum_{i=1}^N \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^N (d_i - \mu)^2 \right) = 0.$$

Solving this equation gives us the value of the estimator for  $\mu$  as

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N d_i.$$

The “hat” notation is commonly used to indicate an estimate. Note that this maximum likelihood (ML) estimate is exactly equal to the mean of the sample. The maximum likelihood





**Figure 2.2:** Maximum likelihood estimation for a Gaussian distribution. **(a)** A set of sample data (black points), the generating distribution (light grey line) and estimated distribution (dark grey line). **(b)** The estimated mean approaches the generative mean as the size of the sample vector increases.

variance also turns out, in this case, to be the given by the (biased) variance of the sample, viz.

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (d_i - \hat{\mu})^2.$$

The two parameters can be estimated separately because  $\hat{\mu}$  has no dependence on  $\hat{\sigma}$ . It is possible to demonstrate, by taking second derivatives, that these estimates really represent a maximum in the likelihood function.

Let's take a step back at this point and consider what we have done. We were given a set of sample values of  $x$ . We hypothesised, and thereafter assumed, that the samples were drawn from a Gaussian distribution with unknown mean and variance. In the language of machine learning, this Gaussian distribution is our **model** for the data, and  $\mu$  and  $\sigma$  are parameters associated with that model. We have no direct way of establishing the values of these parameters, but we used the observed data and Bayes' rule, which can be summarised in words as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}},$$

to learn the most likely estimate for the parameters given the observed data. Since our model describes a distribution which could be used to generate data like  $D$ , it is called a *generative* model.

The process is illustrated by Fig. 2.2(a). A sample of 25 points are shown in black—these were sampled from a Gaussian distribution with mean 3 and variance 1, whose p.d.f. is shown by the lighter curve. The learnt model distribution is the darker curve. It can be seen that the peak of the distribution—the mode, which is equal to the mean for a Gaussian distribution—is slightly offset from that of the generating distribution, and the “broadness” of the curve—which indicates the variance—is slightly less. Nevertheless, the estimated distribution may be considered a satisfactory approximation, and thus useful for predicting the general behaviour of the variable  $x$ . Not surprisingly, increasing the size of the sample vector will produce maximum likelihood estimators that are closer, on average, to the generative parameters—as demonstrated by Fig. 2.2(b). This effect is called the law of large numbers.

It should be remembered that the maximum likelihood method implicitly assumes that the priors in Eq. (2.15) are uninformative. If, on the other hand, meaningful prior information is

available, and we wish to take a more firmly Bayesian approach, we can calculate the maximum of the posterior distribution with the prior distribution incorporated into it. This more general approach to choosing an estimate for the parameters is called the maximum *a posteriori* (MAP) method, and it allows us to influence the parameter estimate based on what we know in advance.

## 2.5 Expectation–Maximisation

Unfortunately, it is quite easy to find cases in which simple maximum likelihood estimation is insufficient to find an estimate for a set of parameters. Consider the two-dimensional, or bivariate, version of the Gaussian distribution described by Eq. (2.13). It is

$$P(x, y | \boldsymbol{\mu}, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x - \mu_x)^2 + (y - \mu_y)^2}{2\sigma^2}\right). \quad (2.16)$$

This is effectively a special case of Eq. (2.14), because we are treating the  $x$  and  $y$  dimensions as independent. This will only be the case if the covariance between  $x$  and  $y$  is zero; but we make that assumption here to avoid overcomplication. Note also that the mean,  $\boldsymbol{\mu} = (\mu_x, \mu_y)$ , is now a vector quantity since it has a component in each dimension. Consider now

$$P(x, y | \theta) = aP_1(x, y | \theta) + (1 - a)P_2(x, y | \theta), \quad (2.17)$$

where each of  $P_1$  and  $P_2$  have the distribution given in Eq. (2.16), and  $\theta = \{\boldsymbol{\mu}_1, \sigma_1, \boldsymbol{\mu}_2, \sigma_2\}$  is a collection of all the parameters of this model. Eq. (2.17) is called a Gaussian *mixture model*, because it is made up of a combination of two independent Gaussian distributions over the same parameter space. The parameter  $a$ , which must be in the interval  $[0, 1]$  to ensure that the overall distribution is properly normalised, is called the mixture coefficient. We include it in the set  $\phi = \{\boldsymbol{\mu}_1, \sigma_1, \boldsymbol{\mu}_2, \sigma_2, a\}$ , a superset of  $\theta$ .

In a generative sense, any sample data point must be drawn from exactly one of the component distributions,  $P_1$  and  $P_2$ . We say there is a **latent variable**, which we denote  $z_i$ , associated with each data point,  $\mathbf{d}_i$ . We can characterise this variable by defining

$$z_i = \begin{cases} 1 & \text{if } \mathbf{d}_i \text{ was drawn from } P_1 \\ 0 & \text{otherwise.} \end{cases} \quad (2.18)$$

By analogy with the maximum likelihood estimation process for a single Gaussian distribution, we might expect to be able to infer the mean and variance of  $P_1$  according to

$$\hat{\boldsymbol{\mu}}_1 = \frac{\sum_{i=1}^N z_i \mathbf{d}_i}{\sum_{i=1}^N z_i} \quad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N z_i \|\mathbf{d}_i - \hat{\boldsymbol{\mu}}_1\|^2}{\sum_{i=1}^N z_i}, \quad (2.19)$$

where  $\|\cdot\|$  is the Euclidean norm; and similarly for  $P_2$ . (Note that  $\sum_i z_i$  is equal to the number of data points that were drawn from  $P_1$ .) However, without any knowledge of the set  $Z = \{z_i\}$ , Eq. (2.19) cannot be evaluated, and so no estimate for  $\phi$  can be calculated. Conversely, if  $\phi$  were known then  $Z$  could be inferred, but we have neither.

The **Expectation–Maximisation** (EM) method provides a way to estimate both  $\phi$  and  $Z$  simultaneously, thus sidestepping the problem of their mutual dependency (Dempster *et al.*, 1977). The method is initialised by choosing a first estimate,  $\hat{\phi}$ , for the parameters. After that, an expectation step, or “E-step”, and a maximisation step, or “M-step”, are applied iteratively until some termination criterion is met. Each E-step calculates a posterior distribution for  $Z$  based on the current parameter estimate, while the M-step updates the parameters.

We once again assume that the elements of our data set,  $D = \{\mathbf{d}_i\}$ , are i.i.d., and hence the values of  $z_i$  are also independent. As a result, the posterior over  $Z$  can be expanded to

$$P(Z | D, \hat{\phi}) = \prod_{i=1}^N P(z_i | \mathbf{d}_i, \hat{\phi}), \quad (2.20)$$

and so we can consider the posterior for each  $z_i$  individually. Bayes' rule gives us

$$P(z_i | \mathbf{d}_i, \hat{\phi}) = \frac{P(\mathbf{d}_i | z_i, \hat{\phi}) P(z_i | \hat{\phi})}{\sum_{z_i} P(\mathbf{d}_i | z_i, \hat{\phi}) P(z_i | \hat{\phi})}, \quad (2.21)$$

where  $\sum_{z_i}$  is shorthand for the sum over the sample space of  $z_i$ . Note that the distributions over  $z_i$  are discrete, so the prior  $P(z_i = 1)$  is meaningful, and will in general be nonzero. Its exact value will be given by the current estimate for the mixture coefficient,  $\hat{a}$ , which is updated by the  $\mathbf{M}$ -step below; and  $P(z_i = 0)$  follows directly by normalisation.

Observe that the particular case  $P(\mathbf{d}_i | z_i = 1, \hat{\phi})$  is equivalent to  $P_1(\mathbf{d}_i | \hat{\theta})$ , a fact that follows straight from the definition of  $z_i$  in Eq. (2.18). As a result, we can expand Eq. (2.21) by exhaustive enumeration of the two outcomes, as follows.

$$P(z_i = 1 | \mathbf{d}_i, \hat{\phi}) = \frac{\hat{a} P_1(\mathbf{d}_i | \hat{\theta})}{\hat{a} P_1(\mathbf{d}_i | \hat{\theta}) + (1 - \hat{a}) P_2(\mathbf{d}_i | \hat{\theta})} \quad (2.22)$$

$$P(z_i = 0 | \mathbf{d}_i, \hat{\phi}) = \frac{(1 - \hat{a}) P_2(\mathbf{d}_i | \hat{\theta})}{\hat{a} P_1(\mathbf{d}_i | \hat{\theta}) + (1 - \hat{a}) P_2(\mathbf{d}_i | \hat{\theta})} \quad (2.23)$$

The job of the  $\mathbf{M}$ -step is to refine our current estimate for  $\hat{\phi}$ . In order to do this, we need concrete values for each  $z_i$ . Since the  $\mathbf{E}$ -step has already calculated posterior distributions for  $z_i$  in Eqs (2.22) and (2.23), we simply take as our  $z_i$  values the expectations of these distributions:

$$\langle z_i \rangle = \sum_{z_i} z_i P(z_i) = P(z_i = 1).$$

Note that due to the nature of the definition of  $z_i$ , this expectation is equal to the value of  $P(z_i = 1)$  calculated in Eq. (2.22). Hence, using these values for  $z_i$ , we can update our estimates for the means and variances of  $P_1$  and  $P_2$  with  $\mathbf{ML}$ , according to Eq. (2.19).

All that remains for the  $\mathbf{M}$ -step is to update  $\hat{a}$ , the remaining element of  $\hat{\phi}$ . Our estimate for this parameter is the expected mean value of the set of latent variables, given by

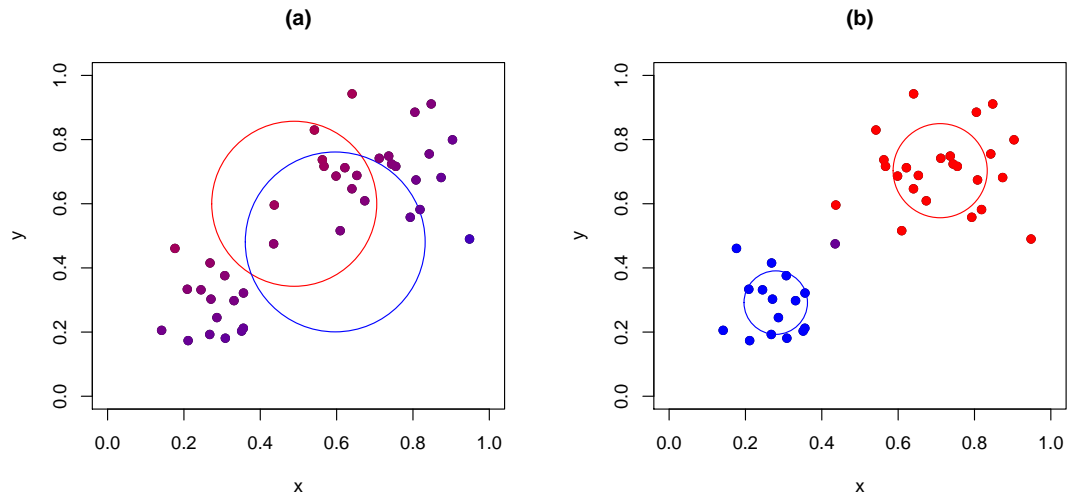
$$\hat{a} = \left\langle \frac{1}{N} \sum_{i=1}^N z_i \right\rangle = \frac{1}{N} \sum_{i=1}^N \langle z_i \rangle = \frac{1}{N} \sum_{i=1}^N P(z_i = 1).$$

Fig. 2.3 shows a graphical representation of the process, in which each small filled circle represents a data point. The posterior distribution over each latent variable, as calculated by the  $\mathbf{E}$ -step, is indicated by a colour, with pure red indicating that  $P(z_i = 1 | \mathbf{d}_i, \hat{\phi}) = 1$ , and pure blue indicating the opposite definite outcome. Hence, the shade of each data point represents how likely it is to be drawn from each of the component distributions. It can be seen that after a single iteration of the algorithm, the estimated component distributions, which are updated by the  $\mathbf{M}$ -step, have a large variance and significant overlap; and as a result the assignment of data to each component is uncertain, so all points appear in shades of purple. By contrast, after 11 further iterations, the algorithm has converged to a stable solution and most points appear red or blue, since they are much more likely to be from one component distribution than the other. There is just one point that remains ambiguous.

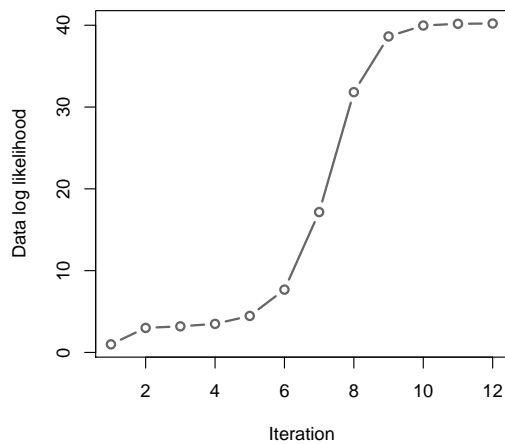
A useful way to gauge the progress of the algorithm is to plot the overall data log-likelihood ( $\mathbf{DLL}$ ), given by

$$\ln P(D | \hat{\phi}) = \sum_{i=1}^N \ln P(\mathbf{d}_i | \hat{\phi}) = \sum_{i=1}^N \ln \left( \sum_{z_i} P(\mathbf{d}_i | z_i, \hat{\phi}) P(z_i | \hat{\phi}) \right),$$

which can be calculated after each iteration of the algorithm. The  $\mathbf{DLL}$  gives us an idea of how well the current model explains the data. Since  $\mathbf{EM}$  is a maximum likelihood technique—differing practically from the simpler  $\mathbf{ML}$  estimation of §2.4 in that it can cope with models that



**Figure 2.3:** Results of applying Expectation–Maximisation to a Gaussian mixture model, after one iteration (a) and at convergence (b). Each large circle represents a component distribution, centred at the mean and with radius equal to one standard deviation. Data points with  $z_i$  closer to 1 are more red, and those closer to 0 are more blue. The generating distribution has parameters  $\mu_1 = (0.3, 0.3)$ ,  $\mu_2 = (0.7, 0.7)$ ,  $\sigma_1 = \sigma_2 = 0.1$ , and  $a = 0.5$ .



**Figure 2.4:** Typical plot of data log-likelihood as the Expectation–Maximisation algorithm progresses.

include latent variables—we might expect that the  $\text{DLL}$  would be at its peak when the algorithm terminates.

An example plot of  $\text{DLL}$  is shown in Fig. 2.4. The first  $\text{M}$ -step produces a very large increase in  $\text{DLL}$  (not shown), after which there is a general increase, ending with a final asymptotic convergence on a maximum likelihood value. Note is that there is *never* a drop in  $\text{DLL}$  from one iteration to the next. This is guaranteed by the theory of the  $\text{EM}$  method, which is beyond our scope here (see Bishop, 2006).

## 2.6 Sampling methods

Up to this point we have dealt with very simple, analytically tractable model distributions; and moreover we have been happy to work with a single estimate for the parameters of the model. However, a maximum likelihood estimator for the parameters does not always exist; and in practice it is often useful to be able to fully characterise a distribution over the model parameter space—that is, the joint sample space of all parameters.

Consider a general case in which we have a scalar valued quantity,  $x$ , modelled by a distribution with parameter set  $\theta$ . The now-familiar Bayes' rule defines the posterior distribution for the parameter set according to

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}, \quad (2.24)$$

where

$$P(x) = \int_{\mathcal{A}_\theta} P(x|\theta)P(\theta)d\theta. \quad (2.25)$$

If we can evaluate the normalisation constant, Eq. (2.25), analytically then it will be possible to characterise Eq. (2.24) exactly. The full posterior distribution over  $\theta$  would then be able to provide information on not only the most likely value of  $\theta$ —i.e. the mode of the distribution—but also on the extent to which such an estimate is likely to be valid or useful. For example, the distribution might have multiple modes, in which case taking a single estimate for the parameters may be inappropriate.

The problem is that for a complicated likelihood function, the integral in Eq. (2.25) may be impossible to evaluate analytically, putting exact marginalisation out of reach. Similar problems occur when trying to find the expectation of a function with respect to a complex distribution. In such cases, it may instead be practical to approximately infer the *target density* over  $\theta$  by drawing samples from it. Given a set of these samples,  $\{\theta^{(i)}\}$  for  $i \in \{1..N\}$ , the approximation is then a probability mass function of the form

$$\hat{P}(\theta) = \frac{1}{N} \sum_{i=1}^N P_\delta(\theta|\theta^{(i)}), \quad (2.26)$$

where  $P_\delta(\theta)$  is a p.m.f. analogue of the Dirac delta function:

$$P_\delta(\theta|\theta^{(i)}) = \begin{cases} 1 & \text{if } \theta = \theta^{(i)} \\ 0 & \text{otherwise.} \end{cases}$$

This is the principle of so-called Monte Carlo ( $\text{MC}$ ) methods, which include the sampling techniques described below (for a review see Andrieu *et al.*, 2003). Of course, the approach presupposes that it is possible to evaluate the distribution of interest, but this is the case often enough for the assumption to be tenable for a wide range of practical problems. In fact, it is sufficient to evaluate the target density to within a multiplicative constant, since the approximating p.m.f., Eq. (2.26), is self-normalising. This is extremely useful, because it obviates the need to evaluate the evidence term in Eq. (2.24) when sampling from the posterior distribution.

Moreover, by the law of large numbers the expectation of some function,  $f$ , with respect to  $\hat{P}(\theta)$  will converge towards the expectation of the same function with respect to the true, continuous distribution for  $\theta$  as  $N$  increases:

$$\langle f(\theta) \rangle_{\hat{P}(\theta)} = \frac{1}{N} \sum_{i=1}^N f(\theta) P_{\delta}(\theta | \theta^{(i)}) = \frac{1}{N} \sum_{i=1}^N f(\theta^{(i)}) \xrightarrow{N \rightarrow \infty} \langle f(\theta) \rangle_{P(\theta)} = \int_{\mathcal{A}_{\theta}} f(\theta) P(\theta) d\theta.$$

The issue now becomes one of choosing samples: how can we efficiently generate pseudorandom numbers which accurately represent the unknown target distribution? We are generally primarily interested in regions of the parameter space in which  $P(\theta)$  is relatively large, but how can we identify such places without evaluating the distribution everywhere?

The naïve method of sampling at every point on a grid throughout the space will quickly become unfeasible, especially if the space has high dimensionality—that is, if there are a large number of parameters. The next most simple approach is to choose points randomly and uniformly from the parameter space, and sample the distribution at those points. However, since areas of high probability density are usually concentrated in a small region of the space, the number of samples required to ensure that this *typical set* is reached at least a few times will still often be prohibitively large.

### 2.6.1 Rejection sampling

A more sophisticated general approach to the sampling problem is to avoid sampling directly from the unknown target density,  $P(x)$ , and instead sample from a known, simpler *proposal density*. In particular, if we can evaluate  $\tilde{P}(x) = zP(x)$ , where  $z$  is an unknown constant, and we can find a proposal density,  $Q(x)$ , and a finite positive real number,  $k$ , such that  $\tilde{P}(x) \leq kQ(x)$  for all real  $x$ , then we can apply a method known as *rejection sampling*.

Fig. 2.5(a) shows a situation in which this approach is appropriate. In this case the target density is a Gaussian mixture with component means at  $x = 3$  and  $x = 5$ ; and the proposal density is a simple Gaussian distribution, centred at  $x = 3.5$ , with  $k = 2$ . In a one-dimensional case such as this, it is easy to see by inspection that the proposal density is always greater than the target density.

The process for generating  $N$  samples from the target density is given by Algorithm 2.1. In common with most mc methods, the rejection sampling algorithm involves the use of (uniformly distributed) random numbers. At each step, a candidate sample,  $x^*$ , is generated from the proposal distribution and a random number,  $u$ , is drawn from a uniform distribution over  $[0,1]$ . Then, if

$$u < \frac{\tilde{P}(x^*)}{kQ(x^*)},$$

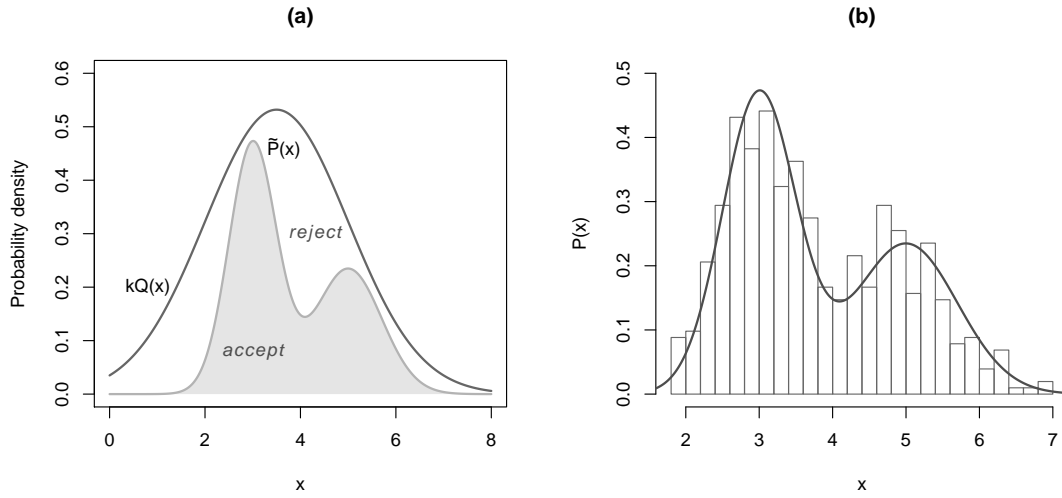
the sample is “accepted” as a sample from  $\tilde{P}(x)$ ; otherwise it is rejected and another candidate sample is drawn. The significance of this acceptance criterion is shown by Fig. 2.5(a): it amounts to a test of whether the quantity  $ukQ(x^*)$ , which is uniformly distributed between

**Require:**  $k \in (0, \infty)$

```

1:  $i \leftarrow 0$ 
2: repeat
3:   Sample  $x^* \sim Q(x)$  and  $u \sim U(0,1)$ 
4:   if  $ukQ(x^*) < \tilde{P}(x^*)$  then
5:      $i \leftarrow i + 1$ 
6:      $x^{(i)} \leftarrow x^*$  [Accept  $x^*$ ]
7:   else
8:     Reject  $x^*$ 
9:   end if
10: until  $i = N$ 
```

**Algorithm 2.1:** Rejection sampling for  $N$  samples.



**Figure 2.5:** Rejection sampling for a univariate Gaussian mixture. **(a)** The target and proposal densities. Samples from the proposal density will be accepted if  $ukQ(x^*) < \tilde{P}(x^*)$ —this corresponds to the shaded area under the target curve. **(b)** Histogram of the accepted samples, overlaid with the exact target density. In this case 51% of samples from the proposal density were accepted.

zero and the value of the proposal density at  $x = x^*$ , falls below the target density. Thus more samples will be accepted in regions where the two densities are very similar, and far fewer in areas where  $\tilde{P}(x) \ll kQ(x)$ . As a result, the technique is most efficient when the proposal density closely approximates the target density. In particular, the two should have as large an overlap in their typical sets as possible. This is certainly the case in our example: both densities are defined for all real numbers, but the vast majority of the probability mass is in the interval  $[0, 8]$ . A uniform proposal density is the worst case, in which case rejection sampling is equivalent to uniform sampling.

After choosing 1000 samples from the proposal distribution, of which 51% were accepted, Fig. 2.5(b) shows a histogram of the accepted samples for a single run of our example case. It can be seen that the normalised histogram agrees quite well with the true target distribution, which is overlaid.

The probability that any given candidate sample is accepted is given by the expectation of the density ratio with respect to the proposal distribution:

$$\Pr(\text{accepted}) = \int_{\mathcal{A}_x} \frac{\tilde{P}(x)}{kQ(x)} Q(x) dx = \frac{1}{k} \int_{\mathcal{A}_x} \tilde{P}(x) dx = \frac{z}{k}.$$

Hence in the example, where  $z = 1$  and  $k = 2$ , we expect around half of samples to be accepted. However, this relationship highlights a crucial shortcoming of rejection sampling—as  $k$  increases, fewer and fewer samples will be accepted, so the run time required to obtain a reasonable sample size from the target density will also increase. For target distributions over high-dimensional sample spaces, it may be hard to find an appropriate value for  $k$  at all; but even if one can be found it will tend to be large, making the method impractical. In such cases, it will be necessary to be more clever about the choice of sampling locations.

## 2.6.2 Markov chain Monte Carlo

A *Markov chain* is a particular type of discrete-time stochastic process in which the state of the system at time  $t$  is dependent only on its state at the previous time step,  $t - 1$ . That is,

$$P(x(t)|x(t-1), x(t-2), \dots, x(0)) = P(x(t)|x(t-1)); \quad (2.27)$$

```

1: Initialise  $x^{(0)}$ 
2: for  $i \in \{1..N\}$  do
3:   Sample  $x^* \sim Q(x|x^{(i-1)})$  and  $u \sim U(0,1)$ 
4:   if  $u < A(x^*, x^{(i-1)})$  then
5:      $x^{(i)} \leftarrow x^*$ 
6:   else
7:      $x^{(i)} \leftarrow x^{(i-1)}$ 
8:   end if
9: end for

```

**Algorithm 2.2:** The Metropolis and Metropolis–Hastings algorithms. The difference between the two methods is in the choice of acceptance function,  $A$ .

the so-called Markov property. The distribution on the right hand side of Eq. (2.27) is called a *transition kernel*.

A subclass of MC techniques called **Markov chain Monte Carlo** (MCMC) methods are designed such that the set of samples drawn forms a Markov chain with the target density as an invariant distribution. Details on how this is achieved can be found in more complete treatments of MCMC methods, such as Neal (1993).

The *Metropolis algorithm* (Metropolis *et al.*, 1953) is an early MCMC method which assumes that the proposal density from which candidate samples,  $x^*$ , are sampled is symmetric in the sense that

$$Q(x^*|x^{(i)}) = Q(x^{(i)}|x^*).$$

Under these circumstances a candidate sample drawn from this distribution is accepted with probability

$$A(x^*, x^{(i)}) = \min \left\{ 1, \frac{\tilde{P}(x^*)}{\tilde{P}(x^{(i)})} \right\}, \quad (2.28)$$

where  $\tilde{P}(x)$  is proportional to the target density,  $P(x)$ , as before. If the candidate sample is accepted then it becomes the new sample,  $x^{(i)}$ ; if not, then the new sample is *the same* as the previous one:  $x^{(i)} = x^{(i-1)}$ . Thus the effect of rejecting a sample differs from the rejection sampling approach in that a new sample is *always* created on each step of the algorithm.

It can be seen directly from Eq. (2.28) that if the value of the target density at  $x^*$  is greater than that at  $x^{(i)}$ , then the sample will always be accepted. On the other hand, if the proposed new sample location represents a substantial drop in probability density, then it is very unlikely to be accepted, and the chain is most likely to remain in its previous state. The result of this policy is that the chain will spend most time in regions of the sample space where the target density is high-valued, as we require.

The Metropolis algorithm was later generalised by W. Keith Hastings to include the case in which the proposal distribution is not symmetric (Hastings, 1970). In this case the acceptance probability is given by

$$A(x^*, x^{(i)}) = \min \left\{ 1, \frac{\tilde{P}(x^*)Q(x^{(i)}|x^*)}{\tilde{P}(x^{(i)})Q(x^*|x^{(i)})} \right\}. \quad (2.29)$$

Algorithm 2.2 describes the Metropolis and Metropolis–Hastings algorithms, given appropriate forms for  $A$ . It is important to note that unlike the rejection sampling method,

```

1: Initialise  $\mathbf{x}^{(0)}$ 
2: for  $i \in \{1..N\}$  do
3:   Sample  $x_1^{(i)} \sim P(x_1|x_2^{(i-1)}, x_3^{(i-1)}, \dots, x_n^{(i-1)})$ 
4:   Sample  $x_2^{(i)} \sim P(x_2|x_1^{(i-1)}, x_3^{(i-1)}, \dots, x_n^{(i-1)})$ 
5:   etc.
6: end for

```

**Algorithm 2.3:** Gibbs sampling over a vector quantity,  $\mathbf{x}$ .



Metropolis–Hastings generates correlated, rather than independent, samples. However, if a subset consisting of, say, every 50th sample is taken, then these may be considered to be close enough to independent for most practical purposes. The proportion of samples which may be kept whilst retaining approximate independence will depend on the exact form of the proposal density, as will the performance of the method in approximating its target. In particular, if the variance of the proposal density is very large, few candidate samples will be accepted, resulting in highly correlated samples; and if it is very small then some significant regions of the parameter space may be left unexplored.

The extension of these methods to the multivariate case where each sample is a vector,  $\mathbf{x}^{(i)}$ , just requires that the proposal distribution be defined in the appropriate number of dimensions. There is no change needed to the algorithms themselves. However, under a popular special case of the Metropolis–Hastings algorithm called *Gibbs sampling*, each element of such a vector is sampled from a different proposal distribution (Geman & Geman, 1984). This method requires that the conditional distributions of each element in the sample vector given all other elements be known, because these are used as the proposal distributions (see Algorithm 2.3). It can be shown that under these circumstances, the acceptance probability for samples is unity, and so this method is highly efficient.

## 2.7 Summary

In this chapter we have reviewed the basic principles of probability, and explained how the strict, frequentist interpretation of probability can be broadened to encompass any proposition with which uncertainty is associated. We have also looked at the basic mechanisms of inference and learning from data, which typically involve the use of Bayes' rule. The rationale for maximum likelihood and maximum *a posteriori* parameter estimates has been explained, and methods for calculating such estimates, including the Expectation–Maximisation approach, have been outlined. Finally, we explored ways in which a probability distribution, whose exact form cannot be calculated analytically, can be approximated efficiently from data. The probabilistic perspective will appear commonly throughout the remainder of this thesis, and we will outline techniques which rely on some of the tools and ideas marshalled above.



## 3

# Magnetisation, excitation and relaxation

AT A FUNDAMENTAL level, the NMR effect is reliant on spin, a phenomenon which, on the scale of individual nuclei, is firmly in the realm of quantum mechanics. Population differences in spin state amongst a very large number of these nuclei give rise to a residual magnetisation, which in turn is the reason that we can retrieve a signal during an MRI scan. By adding energy to a stable system of spins, we can provoke a change in the magnetisation pattern of the system, which can be measured as the spins relax back to their resting states. Moreover, by applying sequences of excitations to brain tissue, and fine-tuning the relaxation process, images of the tissue can be recovered. This chapter provides a basic grounding in these processes, to support the material that follows in later chapters.

### 3.1 State and spin

For a simple Newtonian system such as a moving ball, the dynamical state of the system consists of such quantities as position and momentum, which can in principle be established exactly, and which describe the instantaneous behaviour of the ball with certainty. In a quantum mechanical system, on the other hand, dynamical variables such as position do not have well-defined values at any given time; instead, quantum mechanical theory describes probability distributions over these variables. A measurement of position, for example, is therefore a nondeterministic experiment; and until such a measurement is made, the state of any single quantum object is uncertain.

A form of notation introduced by Paul Dirac allows us to discuss quantum state in abstract terms without concerning ourselves with the details of the particular system we are working with. Using this *bra-ket notation*, quantum state can be described and manipulated using the familiar principles of linear algebra (Dirac, 1958). Full details of the underlying physics, as well as a far more detailed general introduction to quantum mechanics than the sketch which follows, can be found in Bransden & Joachain (1989).

Under Dirac's system, the instantaneous state of a quantum mechanical system is represented by a vector in some state space over the complex numbers, whose dimensionality depends on the characteristics of state in which we are interested. These vector elements of the state space are known as ket vectors, or **kets**, and are written using the notation  $|\cdot\rangle$ , where the dot is to be replaced by a label. The formulation is such that the direction of these vectors is the only property that distinguishes one state from another; lengths are immaterial, and so generally normalised. Consequently,  $|\psi\rangle = c|\psi\rangle$  for any nonzero complex scalar,  $c$ . On the other hand, some combination

$$|x\rangle = x_1|\psi_1\rangle + x_2|\psi_2\rangle$$

is, in general, different to each of the states  $|\psi_1\rangle$  and  $|\psi_2\rangle$ . In fact, the composite ket,  $|x\rangle$ , represents a *superposition* of the two constituent states. The significance of this will be explained

shortly.

If we assume that some set of ket vectors,  $\{|\psi_i\rangle\}$ , forms a basis for the state space we are interested in, then any arbitrary ket can be represented as some linear combination of the set, whose coefficients form a column vector (i.e. single-column matrix):

$$|x\rangle = \sum_{i=1}^n x_i |\psi_i\rangle = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \Psi. \quad (3.1)$$

The matrix  $\Psi$  represents the whole basis set. We note briefly that every ket has a corresponding *bra*, denoted  $\langle \cdot |$ , which is formed by taking the adjoint of the ket vector, which is the combined operation of matrix transposition and complex conjugation. Thus, a bra in matrix representation is a row vector whose coefficients are the complex conjugates of the elements of the ket. That is,

$$\langle x| = |x\rangle^\dagger = \Psi^\dagger \begin{bmatrix} x_1^* & x_2^* & \dots & x_n^* \end{bmatrix},$$

where  $^\dagger$  represents the adjoint, and  $^*$  the conjugate. By multiplying together a bra and a ket, we obtain

$$\langle x|y\rangle = \sum_{i=1}^n x_i^* y_i \langle \psi_i | \psi_i \rangle,$$

which simplifies to

$$\langle x|y\rangle = \sum_{i=1}^n x_i^* y_i \quad (3.2)$$

because the basis kets, like all state kets, are normalised to unit length. Eq. (3.2) is exactly the form of the inner product between  $|x\rangle$  and  $|y\rangle$ .

That quantum state spaces are complex-valued is significant. Recall that the complex number  $z = a + ib$  can be written in an alternative polar form,  $z = re^{i\theta}$ , such that

$$a = r \cos \theta \quad b = r \sin \theta \quad r = |z| = \sqrt{a^2 + b^2}$$

and  $i$  is the imaginary unit, with  $i^2 = -1$ . The complex conjugate is then given by

$$z^* = a - ib = re^{-i\theta}.$$

In polar form,  $r$  is sometimes referred to as the amplitude, and  $\theta$  as the phase. It is precisely the fact that quantum theory allows for phase effects which enables it to explain results such as Claus Jönsson's double slit electron diffraction experiment, which demonstrated wavelike behaviour in particles just as Thomas Young had done for light more than a century and a half before (Jönsson, 1974).

It might be expected in such a system as this, where vector length has no physical implication for the state represented by a particular ket, that eigenvectors are of significant importance; and indeed they are central to quantum physics. Physical properties of quantum systems, such as momentum or position, are associated with linear operators in the Dirac formalism. In particular, these so-called *observable* operators are self-adjoint, so that  $A^\dagger = A$ ; and as such their eigenvectors are orthogonal and their eigenvalues are always real (Riley *et al.*, 2002, §8.13.2). As a result the *eigenstates*,  $|\epsilon\rangle$ , which satisfy

$$A|\epsilon\rangle = \lambda|\epsilon\rangle$$

for real scalar values of  $\lambda$ , make up a natural orthonormal basis set for the state space in which the observable  $A$  operates.

One such observable property is **spin**, a quantum characteristic which is intrinsic to particles such as protons and has no classical equivalent. These particles can be thought of as having a natural angular momentum which causes them to spontaneously spin in place. Consider a

single component of this three-dimensional spin, along a direction which we will choose to be the  $z$  axis of some physical space—in the case of protons, which are abundant in brain tissue, the corresponding spin operator,  $S_z$ , has two eigenstates, which are called “spin up” and “spin down” and may be thought of as analogous to clockwise and anticlockwise. The *magnetic quantum number* of the proton,  $m$ , takes the value  $\frac{1}{2}$  for the spin up state, and  $-\frac{1}{2}$  for the spin down. Since the eigenstates are orthonormal, the inner product of any pair of them is given by the Kronecker delta. That is,

$$\langle m | m' \rangle = \delta_{mm'} = \begin{cases} 1 & \text{for } m = m' \\ 0 & \text{for } m \neq m' \end{cases} . \quad (3.3)$$

As described in Eq. (3.1), an arbitrary spin state,  $|\psi\rangle$ , can then be described as a linear combination of the spin up and spin down eigenstates:

$$|\psi\rangle = \sum_m p_m |m\rangle .$$

In these circumstances, where the basis vectors are a set of eigenstates, the coefficients,  $p_m$ , are called *probability amplitudes*, and have a specific practical significance: their squared moduli represent the probability masses associated with each basis vector in the state  $|\psi\rangle$ . This probability mass function associated with the state of the system manifests itself when the state is measured, such that

$$\Pr(M = m) = |p_m|^2 = p_m^* p_m , \quad (3.4)$$

where  $M$  is the random variable representing the measured spin value.<sup>a</sup> It is important to remember that a measurement of the spin of a proton can *only* yield one of the two values  $\pm\frac{1}{2}$ , which make up the discrete sample space of  $M$ .

Given this interpretation of the superposed state, we can immediately write down the expected value of a spin measurement by referring back to Eq. (2.9). It is

$$\langle M \rangle = \sum_m m P(m) = \sum_m |p_m|^2 m . \quad (3.5)$$

The orthonormality of the eigenstates described by Eq. (3.3) allows us to expand this as

$$\sum_m \sum_{m'} p_m^* p_{m'} m \langle m | m' \rangle ,$$

which can be expanded in matrix form—provided that  $S_z$  is correctly constructed—to  $\langle \psi | S_z | \psi \rangle$ , a full *bracket*, which is the way that expectation values are written in Dirac notation. Given an obvious formulation of orthonormal eigenstates in this two-dimensional state space, viz.

$$\left| \frac{1}{2} \right\rangle \text{ or } |\uparrow\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \left| -\frac{1}{2} \right\rangle \text{ or } |\downarrow\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix} ,$$

this formulation works out correctly if we take as the spin operator

$$S_z = \frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} , \quad (3.6)$$

which has no off-diagonal components and is therefore trivially self-adjoint.

---

<sup>a</sup>The process of measurement is a crucial and counterintuitively complex one in quantum mechanics. The question of what constitutes a measurement is a controversial one, but the essential outcome is a sampling from the distribution given by Eq. (3.4), and an apparent “collapse” of the system’s state into the eigenstate corresponding to the outcome, so that repeated measurements will all produce this same outcome.

### 3.2 Protons in a magnetic field

It is unlikely to come as a surprise that quantum state is not a time-invariant phenomenon. The observable that determines the evolution over time for a quantum mechanical system is energy, which is represented mathematically by a *Hamiltonian* operator,  $H$ . Given the appropriate Hamiltonian, the change in state of the system is described in general by the famous time-dependent *Schrödinger equation*. In the special case where the Hamiltonian itself has no time dependence, the general equation can be separated into two: the time-independent Schrödinger equation, which takes the form of an eigenvalue equation; and a relationship describing the time evolution of the system (Bransden & Joachain, 1989). Specifically, we get

$$H|\psi(t)\rangle = E|\psi(t)\rangle \quad (3.7)$$

and

$$H|\psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle, \quad (3.8)$$

where  $\hbar$  is the reduced Planck constant, which corresponds to the size of a fundamental quantum of energy. Integrating Eq. (3.8) leads directly to the solution

$$|\psi(t)\rangle = \exp\left(\frac{-iHt}{\hbar}\right) |\psi(0)\rangle. \quad (3.9)$$

Now, we may note that if  $|\psi(0)\rangle$  is an eigenstate of the Hamiltonian with eigenvalue  $E$ —as per Eq. (3.7)—then  $H$  will be replaced by  $E$  in this solution, and the time evolution of the system will amount to a mere multiplication of the eigenstate by a complex constant; which, as we know, has no effect on the physical state of the system. Consequently, a system that is in a state that is an eigenstate of the Hamiltonian will stay in that state, unless some external influence dislodges it.

Nuclei with spin, such as that of hydrogen ( $^1\text{H}$ , which contains just a single proton), act like tiny magnets. If all three components of the spin of a proton are represented by  $\mathbf{S}$ , then it will have a magnetic dipole moment of  $\boldsymbol{\mu} = \gamma\hbar\mathbf{S}$ , where  $\gamma$  is called the **gyromagnetic ratio**, which varies from one species of nucleus to another. As a result of this dipole moment, an external magnetic field will have a significant effect on these nuclei. The Hamiltonian corresponding to the interaction with a static magnetic field applied in the  $z$  direction is given by

$$H = -\gamma\hbar B_0 S_z, \quad (3.10)$$

where  $B_0$  is the field strength (Callaghan, 1991). By substituting Eq. (3.10) into Eq. (3.7) and rearranging, we obtain

$$S_z |\psi(t)\rangle = -\left(\frac{E}{\gamma\hbar B_0}\right) |\psi(t)\rangle,$$

which is an eigenvalue equation for  $S_z$ . However, we already know that the eigenvalues of this operator are  $\pm\frac{1}{2}$ , so we can obtain immediately

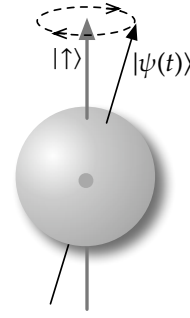
$$-\left(\frac{E}{\gamma\hbar B_0}\right) = \pm\frac{1}{2};$$

and so

$$E = \mp \frac{\gamma\hbar B_0}{2}. \quad (3.11)$$

Notice the signs: the energy of the spin up state is lower than that of the spin down state. It is clear from Eq. (3.11) that the separation between the energy levels corresponding to the two possible values of  $m$  is given by  $\Delta E = \gamma\hbar B_0$ . This difference is called the *Zeeman splitting*, and gives the magnitude of the energy quantum needed to excite a transition from one Zeeman state to the other. The de Broglie relation,  $\Delta E = \hbar\omega$ , tells us that a photon with angular frequency  $\omega = \gamma B_0$ —the so-called **Larmor frequency**—would be able to supply the required energy.

**Figure 3.1:** A spin state that is not an eigenstate will undergo spontaneous precession about the spin up direction, thereby tracing out the pathway indicated here by a dashed line.



We can also consider the time evolution of this system by substituting Eq. (3.10) into Eq. (3.9), which gives

$$|\psi(t)\rangle = \exp(i\gamma B_0 S_z t) |\psi(0)\rangle. \quad (3.12)$$

The time evolution operator in equation Eq. (3.12)—the exponential term—represents a phase rotation of the state about the  $z$  axis by an angle  $\gamma B_0 t$ . Hence, all noneigenstates will precess about the  $z$  axis (i.e. the direction of  $B_0$ ) at the Larmor frequency. This is illustrated in Fig. 3.1.

As a result of the difference in energy between the two eigenstates, the p.m.f. over spin states for any given proton is not uniform at thermal equilibrium. Rather, a large population of spins will be distributed amongst the Zeeman energy levels according to the Boltzmann distribution, viz.

$$n_{\downarrow}/n_{\uparrow} = \exp(-\Delta E/kT) = \exp(-\gamma \hbar B_0/kT), \quad (3.13)$$

where  $n_{\uparrow}$  and  $n_{\downarrow}$  are the number of spins parallel and antiparallel to the magnetic field respectively,  $k$  is Boltzmann's constant, and  $T$  is the absolute temperature. However, in a 1.0 T field and at normal body temperature (310 K), the fractional excess of protons in the low energy state,  $(n_{\uparrow} - n_{\downarrow})/(n_{\uparrow} + n_{\downarrow})$ , is only  $3.295 \times 10^{-6}$ . Nevertheless, this small difference is significant, and in a large population of spins within a some small region of space—known as an *isochromat*—it is large enough to be measurable on a macroscopic scale.

The **magnetisation** of an isochromat at equilibrium, containing a net excess of  $N$  spins in the positive  $z$  direction, is defined as  $\mathbf{M}_0 = N\boldsymbol{\mu}$ , where  $\boldsymbol{\mu}$  is the proton magnetic moment discussed above. It follows from Eq. (3.13), therefore, that the magnitude of this vector is approximately given by

$$M_0 = \chi B_0 \simeq \frac{\gamma^2 \hbar^2 B_0 (n_{\uparrow} + n_{\downarrow})}{4kT}. \quad (3.14)$$

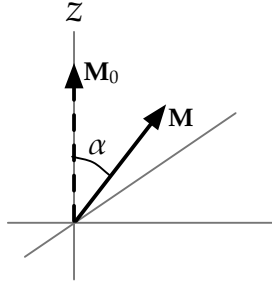
The factor  $\chi$ , which links the magnetisation of the isochromat with the static field strength, is called its *magnetic susceptibility*.

It is clear from Eq. (3.14) that the magnetisation can be increased for a fixed group of spins, thereby increasing sensitivity, by increasing the field strength or decreasing the temperature. However, since the change would have to be substantial to make any significant difference, the latter option is not very practical for *in vivo* MRI!

At rest, the direction of the magnetisation vector,  $\mathbf{M}_0$ , is the same as that of the static field. However, in a nonequilibrium state, this vector could have any arbitrary direction. Whilst the underlying spins must always yield one or other of the eigenstates when measured, the semiclassical representation of spin dipole moments as magnetisation is approximately continuous, since it denotes the aggregate tendency of a large number of individual quantised states.

### 3.3 The NMR signal

As we have seen, a nucleus with spin can be excited from the spin up to the spin down state using electromagnetic radiation with an angular frequency corresponding to the appropriate Larmor frequency. For  $^1\text{H}$  nuclei, this corresponds to a linear frequency of about  $42.5 \text{ MHz T}^{-1}$ , which is in the radiofrequency (RF) range.



**Figure 3.2:** The effect of applying a radiofrequency electromagnetic pulse to a spin isochromat at equilibrium is to “flip” the magnetisation vector,  $\mathbf{M}$ , by an angle  $\alpha$  towards the transverse ( $x$ - $y$ ) plane.

If a spin isochromat is excited so that half of the “excess” protons—which are responsible for the residual magnetisation at equilibrium—are expected to be in each of the spin up and spin down states, then clearly no magnetisation in the longitudinal ( $z$ ) direction remains, because there is no longer any net difference between the populations of spins in each state. However, since the states of the spins are individually precessing about the  $z$  axis, as shown in Fig. 3.1, a net magnetisation in the transverse ( $x$ - $y$ ) plane can arise if the spins are in phase with one another, due to constructive interference. This is exactly the effect of applying an RF pulse to an isochromat at equilibrium (see Fig. 3.2). Since the initial magnetisation vector,  $\mathbf{M}_0$ , does not precess, the spins will be in phase after the excitation has “flipped” the magnetisation towards the transverse plane. The exact angle,  $\alpha$ , by which the magnetisation vector is deflected will depend on the power of the RF pulse and the length of time over which it is applied; but these parameters can be calibrated so as to produce predictably any flip angle required.

The evolution over time of the components of the magnetisation vector due to precession is given by

$$\frac{d\mathbf{M}}{dt} = \gamma \mathbf{M} \wedge \mathbf{B}, \quad (3.15)$$

where  $\wedge$  is the vector cross product, and  $\mathbf{B}$  is the total magnetic field. The latter primarily consists of the static field,  $\mathbf{B}_0$ , but the RF pulse also induces a small and fluctuating field,  $\mathbf{B}_1$ , perpendicular to the longitudinal direction.

Eq. (3.15) does not, however, represent the whole picture. An excited isochromat will not merely precess indefinitely at a fixed angle from the longitudinal direction; rather, its magnetisation will gradually return to the equilibrium state. This relaxation is caused by a combination of processes. Firstly, some of the excitation energy will be spontaneously transferred to the environment as heat—an exponential decay process known as spin-lattice relaxation. Evolution of the  $z$  component of the magnetisation is therefore more accurately reflected by

$$\frac{dM_z}{dt} = \gamma(M_x B_y - M_y B_x) - \frac{M_z - M_0}{T_1}, \quad (3.16)$$

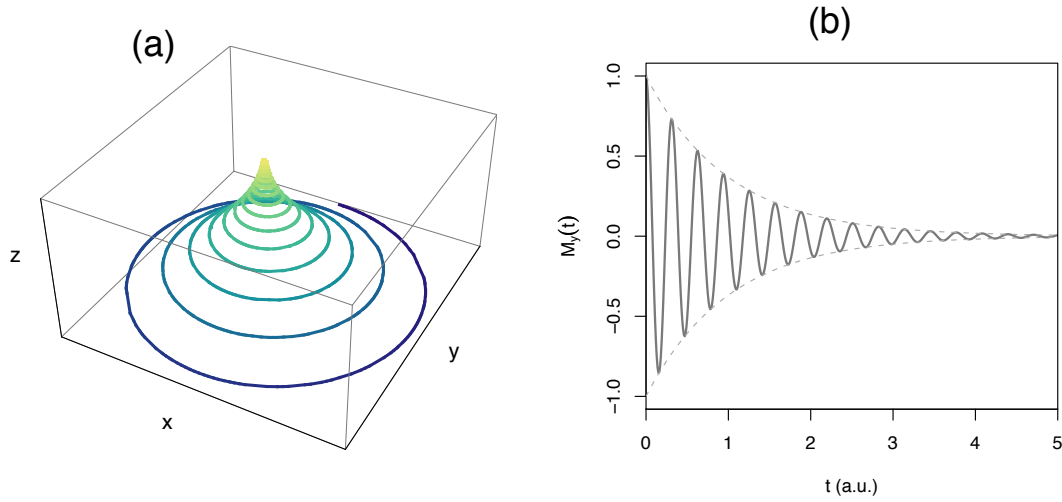
where  $T_1$  is a time constant. The second relaxation process involves the transfer of energy between excited spins, which causes their rates of precession to vary slightly from one to the other. This in turn results in a dephasing of the spin states, so that the transverse component of the magnetisation vector diminishes; again exponentially:

$$\frac{dM_x}{dt} = \gamma(M_y B_z - M_z B_y) - \frac{M_x}{T_2} \quad \frac{dM_y}{dt} = \gamma(M_z B_x - M_x B_z) - \frac{M_y}{T_2}. \quad (3.17)$$

This second time constant  $T_2$  is, in general, not equal to  $T_1$ ; but it cannot be larger. Eqs (3.16) and (3.17) are collectively the *Bloch equations* for nuclear induction (Bloch, 1946).

Once the RF pulse has been applied to excite the system it is switched off, leaving the  $z$  component of  $\mathbf{B}$  as the only nonzero one (the static field is still on). Under these conditions we can therefore ignore all terms in the Bloch equations containing  $B_x$  or  $B_y$ . The resulting





**Figure 3.3:** Effects of relaxation on the magnetisation vector after a  $90^\circ$  excitation pulse. **(a)** The vector precesses around the  $z$  axis with a monotonically decreasing radius. **(b)** The  $y$  component of the relaxing magnetisation vector (or equivalently, the  $x$  component) induces a decaying voltage in the receive coil. In both subfigures,  $T_1 = 2T_2$ .

simplified differential equations can be integrated to give the solutions

$$\left. \begin{aligned} M_x(t) &= (c_1 \cos(\gamma B_z t) + c_2 \sin(\gamma B_z t)) e^{-t/T_2} \\ M_y(t) &= (c_2 \cos(\gamma B_z t) - c_1 \sin(\gamma B_z t)) e^{-t/T_2} \\ M_z(t) &= M_0 + c_3 e^{-t/T_1} \end{aligned} \right\}, \quad (3.18)$$

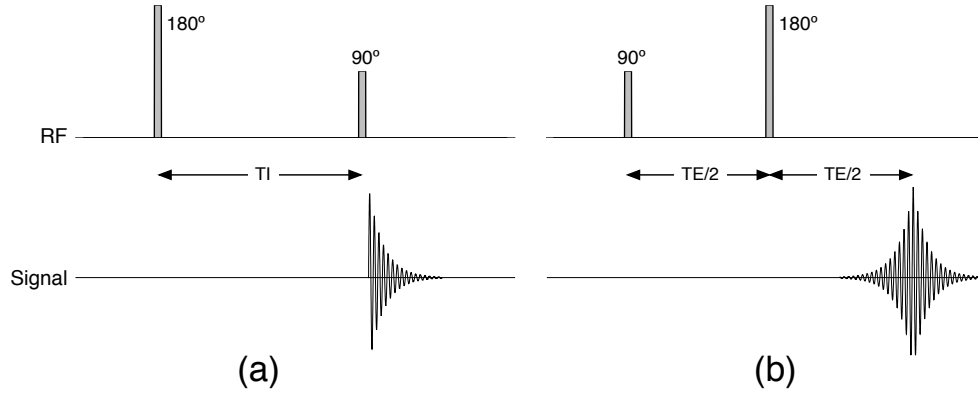
where  $c_1$ ,  $c_2$  and  $c_3$  are constants; although there is no loss of generality in taking  $c_1 = 0$ , so we will do so. The  $x$  and  $y$  components of the magnetisation will then trace out a circle of radius  $c_2 e^{-t/T_2}$  with angular frequency  $\gamma B_z$ , which is the Larmor frequency for the main field. This radius is itself dependent on time, clearly, and will monotonically decrease as relaxation proceeds; as shown in Fig. 3.3(a).

If an electrically conducting coil is placed around the subject in the transverse plane, the rotating transverse magnetisation component will induce a voltage in it—just as in an electrical generator—whose magnitude will decay exponentially due to relaxation (see Fig. 3.3(b)). It is this phenomenon, known as a **free induction decay** (FID), which forms the signal for an NMR experiment. Note that  $M_x$  and  $M_y$  differ only in phase, and they make up the real and imaginary components of the complex-valued oscillating function  $M_{xy}(t) = c_2 e^{-t/T_2} e^{i\omega t}$ , where  $\omega = \gamma B_z$  above. It is often convenient to work with the transverse magnetisation in these terms.

### 3.4 Pulse sequences

The relaxation time constants,  $T_1$  and  $T_2$ , are not invariant throughout the brain; or, indeed, the body (de Certaines *et al.*, 1993). Moreover, there can be systematic differences in these parameters between healthy and pathological tissue of the same basic type. It is therefore constructive from a clinical point of view to devise NMR-based protocols for measuring rates of relaxation; or at least, for creating contrast between regions whose rates differ. This aim can be achieved by applying carefully designed sequences of RF pulses to brain tissue.

A simple pulse sequence for weighting the signal by the value of  $T_1$  is called inversion recovery. At its simplest, this sequence consists of a pulse inducing a flip angle of  $180^\circ$ ,



**Figure 3.4:** Pulse sequence timing diagrams for inversion recovery (a) and spin-echo (b) sequences. The axis represents time, but pulse and signal widths are not to scale.

followed after a time  $\tau_1$  by a  $90^\circ$  pulse. The first of these—the inversion pulse—will flip an isochromat at equilibrium so that all of the magnetisation is antiparallel to the static field. The system will then decay back towards the equilibrium state according to

$$M_z(t) = M_0 \left(1 - 2e^{-t/T_1}\right), \quad (3.19)$$

until the second pulse is applied to convert the remaining longitudinal magnetisation into measurable transverse magnetisation. Note that Eq. (3.19) is a special case of Eq. (3.18) in which  $c_3 = -2M_0$ , the choice that produces the correct boundary conditions. By measuring the FID amplitude for several values of the inversion time,  $\tau_1$ , one can infer the value of  $T_1$  in a sample.

The same pair of pulses in the opposite order can be used to give  $T_2$ -weighting, in a technique known as spin-echo (Hahn, 1950). The spins are allowed to dephase for a time  $T_E/2$ , after which the magnetisation is flipped. After another time period of  $T_E/2$  the spins, which are now dephasing in the opposite sense, will return to being in phase with one another, thus producing a measurable signal. Once again, the constant  $T_2$  can be recovered by repeating the experiment with several values of the echo time,  $T_E$ . A significant benefit of this approach is that the separate dephasing effects of small local variations in the main static field—which are always present to some degree—will cancel out at the time the FID amplitude is measured.<sup>b</sup> The transverse magnetisation component will therefore evolve according to

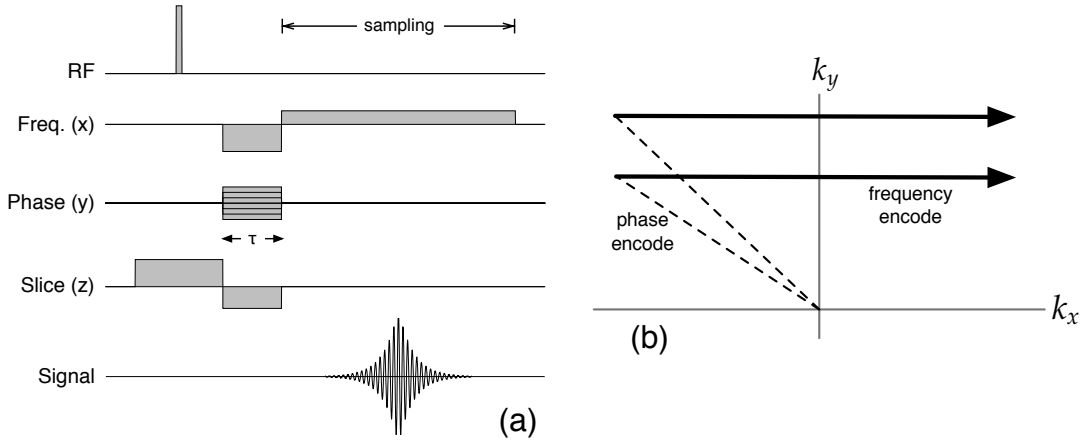
$$M_{xy}(t) = M_0 e^{-t/T_2} e^{i\omega t}, \quad (3.20)$$

a version of Eq. (3.18) with  $c_2 = M_0$ . This is valid as long as *all* of the equilibrium magnetisation is initially flipped into the transverse plane. In order to ensure that this is the case, the repetition time,  $\tau_R$ , between successive  $90^\circ$  pulses in a train of spin-echoes must be sufficiently large to allow the longitudinal magnetisation to recover fully.

The inversion recovery and spin-echo pulse sequences are illustrated in Fig. 3.4, in a schematic representation called a *pulse sequence timing diagram*.

In order to make the move from NMR to MRI, we need the ability to localise a signal in space. Spatial information can be encoded in the signal by applying magnetic *gradients*—that is, static magnetic fields whose strength varies (linearly) across a region of space. The magnitude of these gradients is small compared to that of the main field—typically on the order of  $10^{-2} \text{ T m}^{-1}$ —but they are large enough to provoke variation in the angular frequency at which local magnetisation vectors precess. A gradient with magnitude and orientation described by

<sup>b</sup>In fact, this is only true under the (naïve) assumption that spins do not move during the experiment. In practice, there is movement within the field between the  $90^\circ$  pulse and the signal measurement; a fact which is exploited by diffusion MRI, as we will see in chapter 4.



**Figure 3.5:** The spin-warp imaging sequence. A phase encoding gradient is applied, typically along the  $y$  axis, for a time  $\tau$ ; after which a frequency encoding gradient is applied along the  $x$  axis, and the FID signal is sampled. This process is repeated a number of times with different magnitudes of phase encoding to build up a full three-dimensional brain volume. The sequence is shown as a pulse sequence timing diagram (a) and in terms of its characteristic trajectory in  $\mathbf{k}$ -space (b).

a vector  $\mathbf{G} = (G_x, G_y, G_z)$  will produce a local frequency shift, relative to the Larmor frequency, described by

$$\omega(\mathbf{r}) = \gamma \mathbf{G} \cdot \mathbf{r} = \gamma (G_x r_x + G_y r_y + G_z r_z),$$

where  $\mathbf{r} = (r_x, r_y, r_z)$  represents location in the brain. After applying a  $90^\circ$  RF pulse to create a measurable FID, the signal from a small volume of tissue is therefore given by

$$dA(\mathbf{G}, t) = \rho(\mathbf{r}) \exp(i\gamma t \mathbf{G} \cdot \mathbf{r}) d\mathbf{r}. \quad (3.21)$$

We ignore the effects of spin-spin ( $T_2$ ) relaxation for simplicity, but in a real experiment its effect needs to be quantified. Morris (1986) provides a detailed explanation of the impact it has on the signal. The scalar field  $\rho(\mathbf{r})$  represents the number of spins per unit volume at each location in the brain. This spin density is proportional to the initial magnetisation,  $M_0$ , as we saw in Eq. (3.14); and it is this property that we wish to recover in our experiment. The signal value denoted by the left hand side of Eq. (3.21) is therefore not exactly the FID described by the Bloch equations, but it is closely related to it. Now, introducing

$$\mathbf{k} = \frac{1}{2\pi} \gamma \mathbf{G} t,$$

the signal over the whole brain is given by integrating Eq. (3.21):

$$A(\mathbf{k}) = \int \rho(\mathbf{r}) \exp(i2\pi \mathbf{k} \cdot \mathbf{r}) d\mathbf{r}. \quad (3.22)$$

Eq. (3.22) describes a Fourier relationship between the spin density throughout the brain,  $\rho(\mathbf{r})$ , and the measured signal in the presence of magnetic gradients; and it is therefore the fundamental relationship in MRI. If we sample the signal at a number of locations in  $\mathbf{k}$ -space, we can recover the spin density using a discrete Fourier transform.

There are a number of schemes for traversing  $\mathbf{k}$ -space with various advantages and disadvantages, but we will just describe a relatively straightforward one to give the idea. Fig. 3.5 shows a sequence called spin-warp (Edelstein *et al.*, 1980). It should be noted that this is an imaging sequence using gradients, which is quite independent from the sequences of RF pulses which are used to affect contrast.

The timing diagram in Fig. 3.5(a) shows that after the RF pulse is applied, a gradient is applied for a certain time,  $\tau$ , along the  $y$  axis. The effect is to apply a phase offset to the

magnetisation vectors,  $\gamma\tau G_y r_y$ , which depends on their position in the  $y$  direction—thereby encoding position information in the phase of the signal. Immediately afterwards, another gradient is applied in the  $x$  direction, and is maintained while the signal is sampled. In this case, the frequency of precession of the magnetisation vectors as time progresses is altered by an amount  $\gamma G_x r_x$ —as we saw above—which depends on the location of the tissue along the  $x$  axis. This combination of frequency and phase encoding allows one to spatially locate the source of parts of the signal within a two-dimensional plane. Localisation in the third dimension of space is achieved by selective excitation: that is, only a single “slice” of a certain thickness is excited at a time, and the 3-D image is then built up from a series of these 2-D slices.<sup>c</sup> A slice selection gradient is applied at the same time as the RF pulse.

As a trajectory through  $\mathbf{k}$ -space, the sequence is easily represented. Fig. 3.5(b) shows it in these terms. The phase encode step, along with the application of a negative gradient in the frequency encode direction at the same time, moves us to the “leftmost” position in the space for some value of  $k_y$ . Then, during the application of the frequency encoding gradient, the trajectory moves in the positive  $x$  direction, and all the signal data for this phase encode level is recorded. This process is repeated for several different magnitudes of phase encode gradient, and  $\mathbf{k}$ -space is thereby sampled line-by-line.

The spin-warp sequence requires a separate RF pulse for each line of  $\mathbf{k}$ -space, which limits the rate at which images can be acquired. On the other hand, an influential alternative technique called echo-planar imaging (EPI) is able to reconstruct an entire 2-D slice image using a single excitation pulse or “shot” (Mansfield, 1977). This method is now widely used because of its speed advantages, especially in studies that require a large number of brain volumes to be imaged, such as those using diffusion MRI or functional MRI.

### 3.5 On ghosts and pile-ups

Magnetic resonance images are susceptible to various different types of artefacts, which adversely affect their qualitative and quantitative interpretability and therefore need to be avoided or corrected for whenever possible (see for example Pusey *et al.*, 1986). We describe here the three most significant artefacts for diffusion MRI.

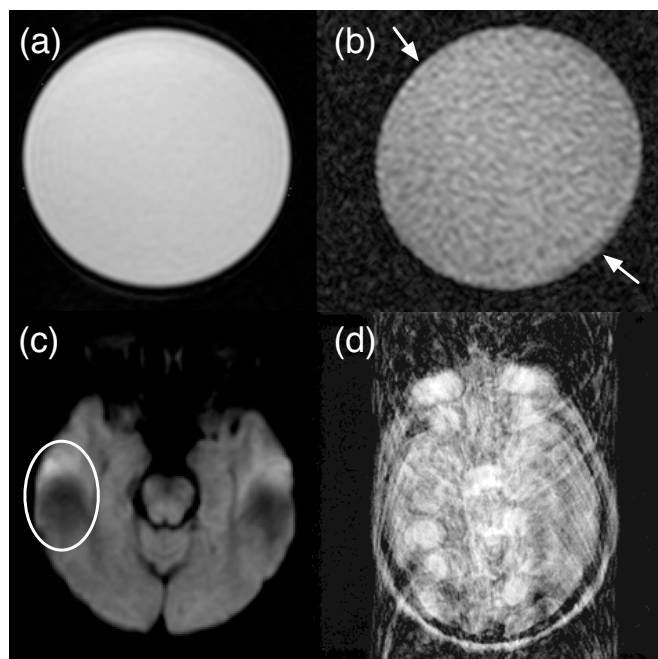
Firstly, there is the problem that the subject, which is usually a living and unsedated human patient or volunteer, may move during the scan. Even if there is no wholesale movement of the head, localised movement can occur as the subject swallows or moves his eyes. The ventricles, which are full of cerebrospinal fluid, typically exhibit spontaneous pulsatile movement; and dilation and contraction of the carotid arteries during the cardiac cycle can also be a source of this kind of artefact. The effect of motion during the sequence is to shift the phase of the signal originating from a particular location, which causes blurring and ghosting—that is, the appearance of nonphysical objects, or of a physical object several times. Motion artefacts can be alleviated by using a sequence that acquires images very quickly—generally EPI—and by “gating” image acquisition so that each slice is collected at the same point in the cardiac cycle (Lanzer *et al.*, 1984).

Whilst EPI is less sensitive to motion effects than other imaging sequences, it is considerably more vulnerable than other methods to two other types of artefact: eddy current induced distortions and susceptibility effects. We will describe these separately.

Eddy currents are tiny circulating electric current loops which are induced by the applied gradient fields, particularly when they are large in magnitude or switched rapidly. These in turn act as electromagnets with magnetic fields that oppose the effect of the gradient field, causing magnification, translation and shearing in the phase encoded direction of the image. The gradients used for diffusion imaging are particularly prone to produce this kind of artefact. One way to significantly reduce their effects is to use a twice-refocussed spin-echo sequence, as described by Reese *et al.* (2003).

Susceptibility effects occur at boundaries between materials with significantly different magnetic susceptibilities—as defined by Eq. (3.14). In the brain this is most obvious near di-

<sup>c</sup>This is the most common arrangement, but it is possible to use phase encoding in two dimensions, in which case selective excitation is unnecessary.



**Figure 3.6:** Examples of various types of MRI artefact. Eddy currents induce a distortion in (b) which results in this circular “phantom” appearing squashed relative to a reference image (a). (Note that the increased noise level in subfigure (b) is not caused by eddy currents.) A susceptibility effect near the ear canals produces signal pile-up and drop-out (c), resulting in artefactual bright and dark patches in the image. Motion by the subject can produce major blurring and ghosting effects (d). Subfigures (a–c) are courtesy of Dr Susana Muñoz Maniega; subfigure (d) is reproduced from Pipe (1999).

viding lines between soft tissue and air—around the sinuses, for example. At such boundaries, the field is locally distorted and therefore rendered inhomogeneous; and as a result signal can “drop out” of some areas while “piling up” in others. Strong susceptibility effects can also be seen if a subject has a small piece of metal near her head, like a hair clip.

Fig. 3.6 illustrates the effects of these different types of artefact. Image (b), which illustrates the distorting effect of eddy currents, is a diffusion-weighted image—as we will see in chapter 4, these images are particularly vulnerable to this sort of artefact. Image (d) is an extreme example of a motion artefact, which makes this image totally unusable.

Whilst not strictly an artefact, there is a further imaging issue which is important when it comes to interpreting MRI data. In practice, the FID is not retained in its original, continuous form, but rather sampled at regular intervals by an analogue-to-digital convertor. As a result the signal in the final image is discretised into spatial units with a fixed volume called **voxels**.<sup>d</sup> The larger the dimensions of these voxels, the higher the signal to noise ratio of the image; but at boundaries between tissue types, the inhomogeneous signal will be averaged across the region represented by the voxel. This implicit averaging is called a **partial volume** effect. These effects make images hard to interpret, since one cannot easily tell what contribution white and grey matter, or healthy and unhealthy tissue, had to the measured signal value.

## 3.6 Summary

Beginning with a single proton, we have described in this chapter how the stochastic behaviour of atomic nuclei can be usefully represented at the macroscopic scale in terms of magnetisation. We have also demonstrated how this phenomenon may be manipulated using radiofrequency radiation, and then measured during relaxation to elucidate characteristics of living tissue. These techniques usually culminate, for clinical purposes, in the creation of images, whose formation we have also discussed. Finally, we have seen that the quality of magnetic resonance images can be affected by a number of artefacts, which arise as side-effects of the scanning process. It should be emphasised that MRI pulse sequence design, and artefact avoidance and correction, are both substantial fields in their own right; and many problems and solutions exist which have not been touched upon at all in the brief coverage of the last two sections.

In the following chapter we will focus on the specific application of the NMR effect to the measurement of diffusion.

<sup>d</sup>The word “voxel” is short for “volume element”, by analogy with “pixel”, which abbreviates “picture element”.



# Diffusion magnetic resonance imaging

THE MAGNETIC resonance application that the rest of this thesis will be concerned with is that of diffusion magnetic resonance imaging (dMRI). This chapter provides a brief description of diffusion and how it can be examined in the brain with dMRI. Mathematical models for diffusion in the brain are also presented, along with their theoretical and practical benefits and limitations. Finally, we look at some of the clinical uses of dMRI.

## 4.1 The Einstein picture

Diffusion is a spontaneous phenomenon in any fluid whose temperature is greater than absolute zero (0 Kelvin). The molecules making up the fluid possess kinetic energy and are therefore constantly moving—the greater the energy, the faster the movement. The direction of this movement is random, and will typically change regularly as molecules collide with one another. Diffusion is often thought of as the process by which concentration gradients are flattened out, and we will initially describe it in these terms; but the principle is equally applicable to the movement of molecules within a fluid composed of a single type of molecule—in the latter case, the process is known as self-diffusion. Diffusion is well described by classical mechanics, so we will not need to make another foray into the quantum domain.

Consider first a one-dimensional example. We denote the concentration of some molecule at location  $x$  and time  $t$  with  $C(x, t)$ . The *flux*, or rate of movement of the molecules normal to the concentration gradient, is then given by

$$F = -D \frac{\partial C}{\partial x}, \quad (4.1)$$

where  $D$  is a constant known as the **diffusivity** of the fluid. As a result of this flux, however, the local concentration gradient will decrease, and so a time-dependent aspect needs to be introduced to describe the picture more fully. The equation

$$\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2} \quad (4.2)$$

was first arrived at by Adolf Fick, and so Eqs (4.1) and (4.2) are called Fick's Laws of diffusion (Fick, 1855; reprinted in translation in Fick, 1995).

If we assume that there are  $n$  molecules in total, all of which are at the location  $x = 0$  at time  $t = 0$ —so that  $C(x, 0)$  is a Dirac  $\delta$ -function—then diffusion will proceed such that

$$C(x, t) = \frac{n}{\sqrt{4\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right), \quad (4.3)$$

as described by Einstein (1905). If we divide Eq. (4.3) by  $n$ , we obtain a properly normalised p.d.f. that describes the distribution of the molecules in the  $x$  dimension—or rather, since the distribution is dependent on  $t$ , a continuous-time stochastic process. In particular, we can see by inspection that the distribution  $P(x|t)$  is a Gaussian distribution with  $\mu = 0$  and  $\sigma^2 = 2Dt$ .

If the diffusion process is *isotropic*, or homogeneous across all orientations, then the generalisation to three dimensions is straightforward. Diffusion collinear with each of the vectors  $\mathbf{i}$ ,  $\mathbf{j}$  and  $\mathbf{k}$ —the orthonormal unit vectors in the  $x$ ,  $y$  and  $z$  directions respectively—is independent, and so the joint distribution is given by

$$P(\mathbf{r}|\mathbf{r}_0, t) = \frac{1}{(4\pi Dt)^{3/2}} \exp\left(-\frac{(\mathbf{r}-\mathbf{r}_0)^2}{4Dt}\right), \quad (4.4)$$

where  $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ , and  $\mathbf{r}_0$  is the initial location of the molecules, which is not assumed to be zero in this general case. The mean of the distribution is now a vector,  $\boldsymbol{\mu} = \mathbf{r}_0 = (x_0, y_0, z_0)$ , while the variance is just as it was before:  $\sigma^2 = 2Dt$ .

The dependence of the description above on a concentration gradient does not present a problem for the case of self-diffusion. The fluid molecules may all be of a single species under these circumstances, but we can mentally label a molecule with initial position  $\mathbf{r}_0$  as being (uniquely) of interest; and thereafter treat it as distinct from the rest of the fluid. The mean-squared distance that this molecule will travel during a diffusion time  $t$  is given by

$$\begin{aligned} \langle |\mathbf{r} - \mathbf{r}_0|^2 \rangle &= \langle (x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 \rangle \\ &= \langle (x - x_0)^2 \rangle + \langle (y - y_0)^2 \rangle + \langle (z - z_0)^2 \rangle, \end{aligned}$$

which is equivalent to the sum of the variances along each dimension, since  $\mu_x = x_0$  and so on. We therefore easily arrive at

$$\langle |\mathbf{r} - \mathbf{r}_0|^2 \rangle = \sigma_x^2 + \sigma_y^2 + \sigma_z^2 = 6Dt. \quad (4.5)$$

Note that this equation for the mean-squared diffusion distance has no dependence on  $\mathbf{r}_0$  since the fluid is assumed to be homogeneous, so that diffusion from all starting locations is statistically identical. Wherever a particular molecule starts, its diffusion distance from that point will be the same on average.

In general, diffusion is not isotropic. In a bowl of water it will be very close to isotropic, but in brain tissues—which contain large amounts of water but also various impermeable or semipermeable structures—diffusivity will vary from one direction to another. The Gaussian displacement distribution at time  $t$  therefore has in general the covariance matrix

$$\Sigma = 2Dt = \begin{bmatrix} 2D_{xx}t & 2D_{xy}t & 2D_{xz}t \\ 2D_{xy}t & 2D_{yy}t & 2D_{yz}t \\ 2D_{xz}t & 2D_{yz}t & 2D_{zz}t \end{bmatrix}, \quad (4.6)$$

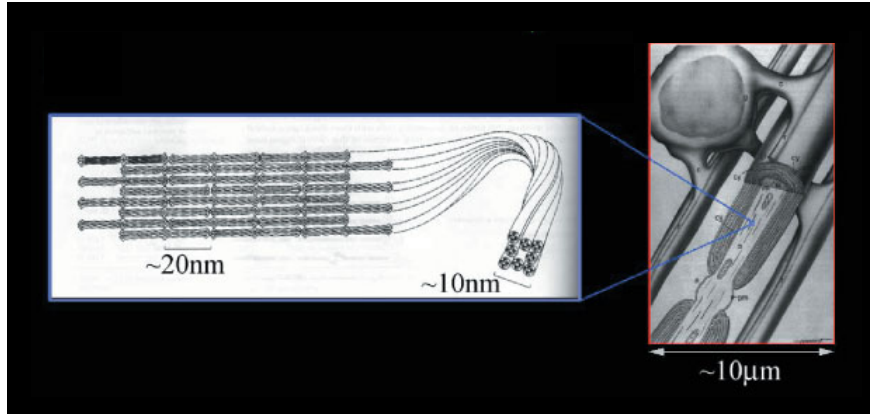
which is symmetric, like any covariance matrix. The diffusivity values making up the matrix  $\mathbf{D}$  are the components of a three-dimensional **diffusion tensor**, relative to the particular orthonormal basis set,  $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ .<sup>a</sup> The special case of isotropic diffusion is then equivalent to the conditions

$$D_{xx} = D_{yy} = D_{zz} = D \quad D_{xy} = D_{xz} = D_{yz} = 0.$$

In the brain, the main diffusing molecular species is water; and since a molecule of water contains two hydrogen nuclei it is visible to MRI. Anisotropic—that is, directionally inhomogeneous—diffusion is associated primarily with white matter, due to the highly linearised structure of this type of tissue (see Fig. 4.1), which is such that the local self-diffusion of water molecules is restricted to a far greater degree across a white matter tract than it is along it. Grey matter, by contrast, lacks any coherent linear structure, and so diffusion around that kind of tissue is much closer to isotropic.

<sup>a</sup>A tensor is an abstract mathematical construction which is independent of the coordinate frame being used. However, relative to any given set of basis vectors, it can simply be represented as a matrix of numbers. Further details would be superfluous here, but can be found in Riley *et al.* (2002).





**Figure 4.1:** The linear microstructure of neural white matter. The axons and glia which make up connective tissue at the micron scale, and the neuronal filaments which are present at the nanometre scale, are mainly collinear, producing a consistent bias in the local self-diffusion of water. Adapted from Mori & van Zijl (2002).

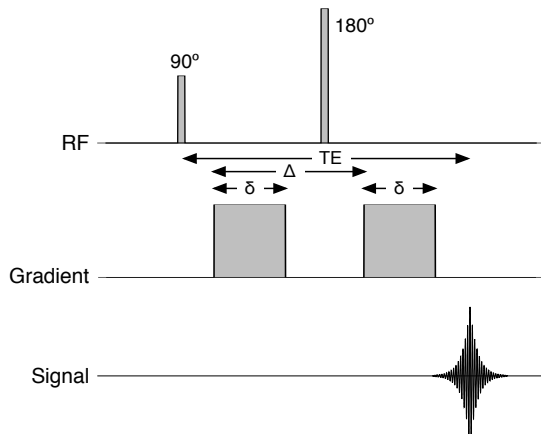
## 4.2 Diffusion tensor imaging

Diffusion sensitisation can be added to the standard spin-echo pulse sequence described in §3.4 by adding a symmetric pair of diffusion weighting gradients either side of the refocussing ( $180^\circ$ ) pulse, as shown in Fig. 4.2. The first of these gradients will offset the phase of the spins by an amount that depends on their location, and the second will provide equal and opposite rephasing *if* the spins have not moved. Since in practice the spins do move, and randomly, an isochromat will become dephased as the component spins spread out. The further the water molecules have diffused during the time,  $\Delta$ , between applications of the gradient, the less perfect this rephasing will be, resulting in a smaller magnitude of final signal. Greater diffusivity is therefore indicated by a more greatly *attenuated* signal. It should be noted that this effect differs from coherent motion or flow, which will produce a phase shift in the spin isochromats, but will not attenuate the signal as random motion does.

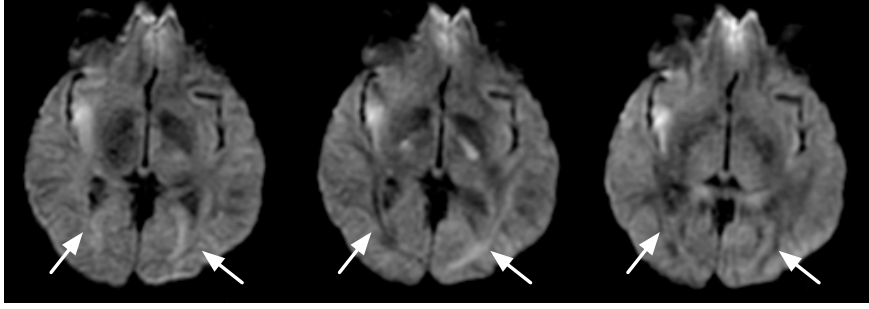
It was shown by Stejskal & Tanner (1965), who first proposed the sequence, that for a diffusion weighting gradient of maximal magnitude  $G$ , applied for a time  $\delta$ , the log-ratio between the signal,  $A$ , after the full echo time,  $T_E$ , and that produced by the initial  $90^\circ$  RF pulse is given by

$$\ln\left(\frac{A(b)}{A_0}\right) = -\gamma^2 \delta^2 \left(\Delta - \frac{\delta}{3}\right) G^2 D^{\text{eff}} = -bD^{\text{eff}}, \quad (4.7)$$

where  $b$ , which incorporates the relevant characteristics of the diffusion gradients, is known as



**Figure 4.2:** Pulse sequence timing diagram for a diffusion-weighted spin-echo experiment. Two diffusion sensitisation gradients are applied either side of the  $180^\circ$  pulse. They are switched on for a time  $\delta$  in each case, and separated by a time  $\Delta$ .



**Figure 4.3:** Diffusion-weighted MR images acquired with diffusion sensitisation along three orthogonal axes. The level of signal attenuation in some areas (such as those indicated with arrows) is evidently dependent on this direction. Images courtesy of Dr Susana Muñoz Maniega.

the diffusion weighting factor—a notation introduced by Le Bihan *et al.* (1986).

It is generally the case that diffusivity appears to vary with time rather than being constant, so Eq. (4.7) describes an *effective* diffusivity,  $D^{\text{eff}}$ , averaged over the diffusion time of the experiment. In tissue with an anisotropic diffusion profile, this “constant” will also vary with the orientation of the diffusion gradient applied to the sample (Moseley *et al.*, 1991; see also Fig. 4.3); and so we need to measure the whole diffusion tensor if we wish to characterise this situation more accurately. The extension of the principles described above to **diffusion tensor imaging** (DTI) was described by Basser *et al.* (1994a). In this case, Eq. (4.7) generalises to

$$\ln\left(\frac{A(\mathbf{b})}{A_0}\right) = -\gamma^2 \delta^2 \left(\Delta - \frac{\delta}{3}\right) G^2 \mathbf{R}^T \mathbf{D}^{\text{eff}} \mathbf{R} = -\sum_i \sum_j b_{ij} D_{ij}^{\text{eff}}, \quad (4.8)$$

where  $\mathbf{R}$  is a normalised column vector describing the direction of the applied gradient, and  $b_{ij}$  are the elements of a symmetric matrix,  $\mathbf{b}$ , which is analogous to the scalar weighting factor in Eq. (4.7). The elements of this weighting matrix encode various interactions between diffusion and imaging gradients, which can be quite complex and which vary from one type of sequence to another (the EPI case is described in Mattiello *et al.*, 1997). The equivalent scalar diffusion weighting factor to  $\mathbf{b}$  is given by the trace of the matrix.

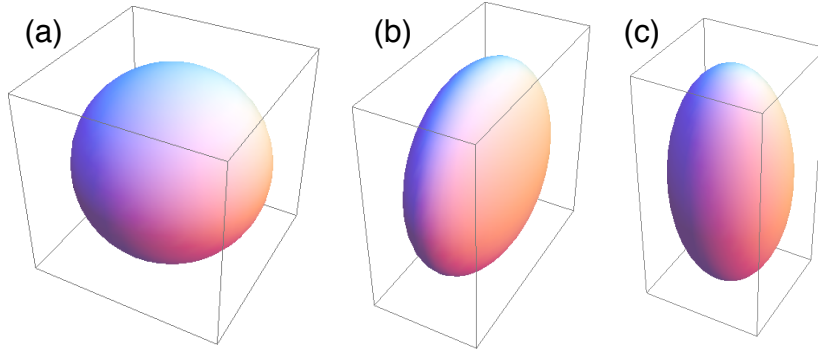
Since knowledge of the pulse sequence design is sufficient to establish the  $\mathbf{b}$  matrix for any given acquisition, Eq. (4.8) represents a system of linear equations that can be solved for the six independent components of the tensor given values of  $A$  for six noncollinear diffusion gradient directions, plus the  $T_2$ -weighted signal,  $A_0$ .<sup>b</sup> However, in practice it is usual to apply more than six different gradient directions, since MR signal measurements are noisy, and then to fit the tensor statistically using multivariate linear regression.

There has been some debate in the literature over the particularities of optimising the choice of gradient scheme for various purposes (Hasan *et al.*, 2001; Papadakis *et al.*, 1999; Skare *et al.*, 2000), particularly the calculation of tensor-derived scalar metrics, which are described below. Broadly speaking, it is as well to acquire data for as many gradient directions as possible (Jones, 2004); and these are commonly arranged to coincide with the vertices of an icosahedron (Batchelor *et al.*, 2003), or to minimise the electrostatic repulsion force when the gradients are treated as point charges<sup>c</sup> (Conturo *et al.*, 1996; Jones, 2004).

Once the effective diffusion tensor has been estimated, it can be used to characterise local diffusion at each voxel in the brain in various ways. Since the matrix representing the tensor is symmetric in any coordinate frame, its eigenvectors are orthogonal and its eigenvalues real. We can therefore construct a local coordinate system from the eigenvectors,  $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ , which are arranged by convention such that the largest eigenvalue is  $\lambda_1$ —corresponding to  $\mathbf{e}_1$ —and the smallest is  $\lambda_3$  (Basser *et al.*, 1994b).

The general shape of the diffusion tensor is commonly visualised using ellipsoids whose radii along each eigenvector direction are given by the square root of the corresponding

<sup>b</sup>Recall from chapter 3 that the basic spin-echo sequence is  $T_2$ -weighted, and this is the only factor in a sequence



**Figure 4.4:** Ellipsoids representing isotropic (a), oblate (b) and prolate (c) diffusion profiles.

eigenvalue (see Fig. 4.4). Thus, the case of isotropic diffusion ( $\lambda_1 = \lambda_2 = \lambda_3$ ) is represented by a sphere, while oblate diffusion ( $\lambda_1 = \lambda_2 > \lambda_3$ ) appears disc-shaped, and prolate diffusion ( $\lambda_1 > \lambda_2 = \lambda_3$ ) appears cigar-shaped.

The average magnitude of the diffusion along the three eigenvectors can be calculated in a rotation-invariant way by taking the trace of the tensor matrix; or equivalently, the sum of the eigenvalues. This quantity is known as the *mean diffusivity* (MD):

$$\text{MD} = \langle D \rangle = \frac{\text{Tr}(D)}{3} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3}. \quad (4.9)$$

This quantity gives no indication of the anisotropy of the tensor, since it takes into account only the mean of the eigenvalues. There is, moreover, no single obvious way to index anisotropy. Three scalar valued measures that have been proposed are fractional anisotropy ( $\text{FA}$ ), relative anisotropy ( $\text{RA}$ ) and the volume ratio ( $\text{VR}$ ), which are defined as follows (cf. Bassar & Pierpaoli, 1996; Pierpaoli & Bassar, 1996).

$$\text{FA} = \sqrt{\frac{3}{2}} \sqrt{\frac{(\lambda_1 - \langle D \rangle)^2 + (\lambda_2 - \langle D \rangle)^2 + (\lambda_3 - \langle D \rangle)^2}{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}} \quad (4.10)$$

$$\text{RA} = \frac{1}{\sqrt{3}} \frac{\sqrt{(\lambda_1 - \langle D \rangle)^2 + (\lambda_2 - \langle D \rangle)^2 + (\lambda_3 - \langle D \rangle)^2}}{\langle D \rangle} \quad (4.11)$$

$$\text{VR} = \frac{\lambda_1 \lambda_2 \lambda_3}{\langle D \rangle^3} \quad (4.12)$$

A VR of unity represents isotropic diffusion, whereas FA and RA are zero when all three eigenvalues are equal. At the other end of the scale, FA and RA are maximal when  $\lambda_2 = \lambda_3 = 0$ , whereas VR is zero when any of the eigenvalues is zero. Of the three, FA gives the highest signal to noise ratio (Papadakis *et al.*, 1999), and is by far the most commonly used in the literature.

### 4.3 A more general displacement distribution

The tensor model makes the assumption that diffusion at the scale of a voxel is essentially Gaussian, which allows us to use the generalised Einstein equation—with covariance matrix given by Eq. (4.6)—as an appropriate model of the underlying process. However this assumption, as we will see later, is not always appropriate; and it is particularly prone to fail in regions where white matter tracts cross one another. Alternative models of diffusion which have been

---

with diffusion gradients of zero magnitude.

<sup>c</sup>The inspiration here is the behaviour of electrons in atomic orbitals, which are equally charged and therefore repel one another. They spontaneously space themselves out as a result.

developed with the particular application of fibre tracking in mind will be discussed in chapter 5, but we will describe here an alternative which predates such applications substantially.

The origins of  $\mathbf{q}$ -space MRI can be traced back to the work of Edward Stejskal, who described how a special case of the Stejskal–Tanner pulse sequence (Fig. 4.2) could be used to infer an arbitrary local displacement distribution. If the time during which the diffusion gradient is applied,  $\delta$ , is made to be very short—in particular so that  $\delta \ll \Delta$ —then the signal attenuation ratio of the experiment is given in general by

$$\frac{A(\mathbf{G}, \Delta)}{A_0} = \int P(\mathbf{r}_0) \int P(\mathbf{r}|\mathbf{r}_0, \Delta) \exp(-i\gamma\delta(\mathbf{r} - \mathbf{r}_0) \cdot \mathbf{G}) d\mathbf{r} d\mathbf{r}_0, \quad (4.13)$$

where the vector  $\mathbf{G}$  embodies the direction and magnitude of the diffusion gradient (Stejskal, 1965). Under the assumption of local homogeneity  $P(\mathbf{r}_0)$ , which represents the initial distribution of diffusing molecules within the volume of interest, is uniform; and so the outer integral can be ignored. Stejskal also showed that if the Gaussian displacement distribution given by Eq. (4.4) were used for  $P(\mathbf{r}|\mathbf{r}_0, \Delta)$ , then Eq. (4.13) becomes equivalent to Eq. (4.8), albeit with  $\Delta - \delta/3$  replaced by  $\Delta$  due to the narrow gradient pulse assumption.

Callaghan *et al.* (1988) later proposed that the direction, magnitude and duration properties of the diffusion gradient be parameterised as

$$\mathbf{q} = \frac{1}{2\pi} \gamma \delta \mathbf{G},$$

by direct analogy with the  $\mathbf{k}$  vector that is so central to magnetic resonance imaging theory (cf. §3.4). Using this notation, and taking  $\mathbf{r}_0 = \mathbf{0}$ , which gives no loss of generality if we assume local homogeneity, the attenuation ratio becomes

$$\frac{A(\mathbf{q}, \Delta)}{A_0} = \int P(\mathbf{r}|\Delta) \exp(-i2\pi\mathbf{q} \cdot \mathbf{r}) d\mathbf{r}, \quad (4.14)$$

which represents a Fourier transform of the displacement distribution. By the Fourier inversion theorem (Riley *et al.*, 2002), we can therefore recover the distribution by means of the inverse transform

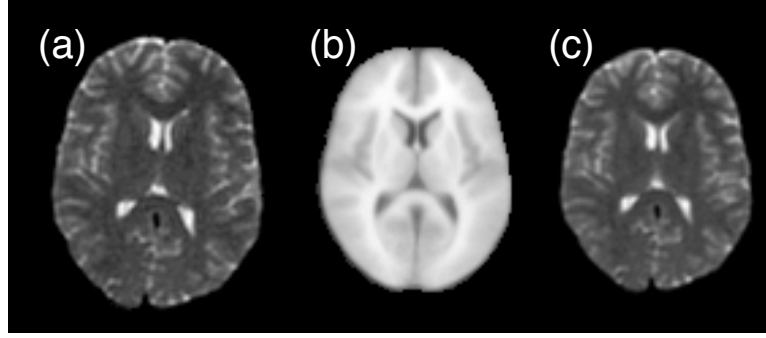
$$P(\mathbf{r}|\Delta) = \frac{1}{A_0} \int A(\mathbf{q}, \Delta) \exp(i2\pi\mathbf{q} \cdot \mathbf{r}) d\mathbf{q}. \quad (4.15)$$

By sampling signal values from a series of locations in  $\mathbf{q}$ -space—typically achieved by incrementally stepping up the gradient strength and changing its direction—one can therefore capture the diffusive behaviour of water molecules in the brain at different length scales and over various diffusion times.

The appeal in acquiring a model-free estimate of the diffusion displacement distribution is clear—modelling assumptions are avoided, and so one need not worry about their validity. However, the narrow gradient pulse assumption made by  $\mathbf{q}$ -space theory is itself problematic. Whilst  $\delta$  must be small enough so that the pulse can be approximated by a Dirac delta function, the time integral of the pulse given by  $\delta\mathbf{G}$  must be finite, otherwise  $\mathbf{q}$  will be zero and there will be no signal attenuation at all. As a result, the magnitude of the gradient pulse needs to be very large. Such gradient strengths are attainable using modern hardware—although they are out of the reach of most clinical MR scanners—but they are very demanding to generate and may have adverse effects on the subject. Hence, studies that have closely approximated the narrow pulse assumption (e.g. Biton *et al.*, 2006, who used the parameters  $\delta = 2$  ms,  $\Delta = 50$  ms,  $G_{\max} = 500$  mT m<sup>-1</sup>) have worked with excised (*ex vivo*) rather than living (*in vivo*) tissue.

## 4.4 The role of registration

Since all but the simplest of dMRI experiments require multiple image acquisitions with different gradient directions, the basic data from which information will be derived is a series of brain volumes. Although motion within volumes will be minimised by using an EPI-based pulse sequence, one cannot rule out the possibility that the subject will move during the whole



**Figure 4.5:** Registration of a  $T_2$ -weighted EPI image (a) to a  $T_1$ -weighted standard brain volume (b) produces a version of the original in standard space (c). Note that the general orientation and scale of subfigure (c) correspond to those of (b), but the details of the image do not match perfectly. The different contrast types of subfigures (a) and (b) is not a problem if the cost function is chosen appropriately.

experiment, particularly if the number of gradient directions is large. Moreover, the varying orientations of the diffusion gradients will result in differing eddy current induced distortion effects from one volume to another. It is therefore unwise to assume that the subject's brain is positioned consistently in the field of view throughout a scanning session.

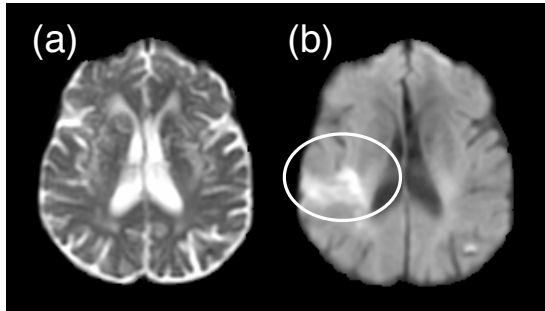
The process of realigning the three-dimensional images is **registration**. Image registration is usually framed as an optimisation problem in which an algorithm attempts to find a global transformation which minimises some cost function indicating the “distance” between two images. A number of cost functions have been used for this purpose, typically based on the correlation or mutual information between image intensity data; but a more divisive issue is the scope of the transformations allowed by the algorithm. The number of degrees of freedom varies from six for a rigid-body transformation—translation by a vector  $\mathbf{L} = (L_x, L_y, L_z)$  and rotation by angles  $\phi$ ,  $\theta$  and  $\psi$  about the  $x$ ,  $y$  and  $z$  axes—up to hundreds or thousands for a complex nonlinear approach, which may involve local as well as global optimisation. Nonlinear methods have the advantage of providing a better match between the original image and the target image, but are slower due to having to optimise over a much larger parameter space, and pose a risk of overfitting.

General purpose linear registration algorithms optimise over affine transformations (Friston *et al.*, 1995; Jenkinson & Smith, 2001; Woods *et al.*, 1998), which incorporate the rigid-body parameters as well as a scaling vector,  $\mathbf{S} = (S_x, S_y, S_z)$ , and three shear terms:  $H_{xy}$ ,  $H_{xz}$  and  $H_{yz}$ . The resulting affine transformation matrix is therefore composed of the product

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 & L_x \\ 0 & 1 & 0 & L_y \\ 0 & 0 & 1 & L_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \phi & \sin \phi & 0 \\ 0 & -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos \theta & 0 & -\sin \theta & 0 \\ 0 & 1 & 0 & 0 \\ \sin \theta & 0 & \cos \theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ \times \begin{bmatrix} \cos \psi & \sin \psi & 0 & 0 \\ -\sin \psi & \cos \psi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & H_{xy} & H_{xz} & 0 \\ 0 & 1 & H_{yz} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} S_x & 0 & 0 & 0 \\ 0 & S_y & 0 & 0 \\ 0 & 0 & S_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

This composite matrix may then be used to transform directly the grid of voxel locations, making up the **native space** in which the original image is acquired, to their equivalent points in the target space. The image data must then be interpolated onto this new grid. The interpolation scheme for this final step may need to be chosen to suit the particular application, but a trilinear scheme is often adequate.

Affine registration of diffusion-weighted images to a  $T_2$ -weighted reference image from the same scanning session is an effective way to correct for eddy current induced distortions in the former, and it simultaneously transforms all of the individual scans into a common space



**Figure 4.6:** Images taken from a patient 11 hours after stroke onset. The  $T_2$ -weighted image (a) is normal in the lesion region, but the averaged diffusion-weighted image (b) shows significantly reduced diffusion compared with the equivalent region in the contralateral hemisphere. Images courtesy of Dr Susana Muñoz Maniega.

so that the correspondence between voxels in each volume is improved. Registration is never perfect, however, and it should be borne in mind that the data used to fit a diffusion tensor (for example) at each voxel cannot truly be said to be taken from a single fixed location in the brain. Some inaccuracy is inevitable.

For comparative studies involving multiple subjects, a popular strategy is to register each subject's reference image to an established standard image such as that described by Evans *et al.* (1993), thus transforming them all into a common **standard space** (see Fig. 4.5). In this case, since no two brains are merely stretched and sheared versions of one another, linear registration is strictly inadequate. The approximation suffices, however, for some purposes.

## 4.5 Diffusion MRI in the clinic

Le Bihan *et al.* (1986) were the first to demonstrate the clinical potential of dMRI. They showed that the presence of astrocytomas (a type of tumour originating in astrocytes) or oedema (swelling due to the accumulation of excess fluid) produced measurable differences in effective diffusivity, when compared with normal tissue. They also demonstrated reduced diffusivity in normal white matter compared to grey matter, which is now established as a standard finding.

Diffusion imaging has been widely used to study acute ischaemic stroke (damage to the brain resulting from a blockage in its blood supply), and has been shown to provide useful information beyond that which is available to structural  $T_1$ - or  $T_2$ -weighted MRI (Baird & Warach, 1998). In particular, reduced diffusivity can be observed in ischaemic tissue very soon after the stroke onset, while  $T_2$  relaxation times are largely unaffected until oedema develops, which takes place much later (Knight *et al.*, 1991; see also Fig. 4.6).

The advent of DTI has made it possible to examine the effects of disease on the diffusion properties of anisotropic tissues—i.e. white matter. With mean diffusivity acting as a proxy for overall water content, and anisotropy indices—in practice, almost invariably FA—indicating the degree of “coherence” or “integrity” of the linear structure intrinsic to white matter, various low level pathological processes such as oedema, neurotoxicity or Wallerian degeneration might plausibly be expected to have some MR-visible impact. These dMRI-derived measures have therefore been applied to investigate the effects of a diverse array of diseases such as multiple sclerosis, amyotrophic lateral sclerosis and Alzheimer's disease (Horsfield & Jones, 2002); as well as psychiatric disorders like schizophrenia, alcoholism and geriatric depression (Lim & Helpert, 2002). More pathologies are being studied year on year.

There has also been significant interest in the effects of normal ageing on white matter (Moseley, 2002; Sullivan & Pfefferbaum, 2006). Anisotropy has been found to be higher in young adults than children (Klingberg *et al.*, 1999), but it then tends to reduce with time through adulthood and into old age (Pfefferbaum *et al.*, 2000), presumably representing the processes of maturation and then degradation of connective tissue. The gradual decline in white matter integrity is accompanied by a tendency for subjects' performance on mental tasks, particularly those using working memory, to decrease with time; and may represent its cause. Moreover, it has been shown that statistically compensating for mental ability in childhood—as measured with an IQ test at age 11—substantially attenuates the relationship, at age 83, between FA and cognitive test performance (Deary *et al.*, 2006), suggesting that childhood IQ may have a bearing on white matter integrity later in life.

Another interesting aspect of normal ageing which has been investigated with dMRI is the phenomenon of leukoaraiosis—also known, rather less concisely, as periventricular white matter hyperintensity—which manifests itself as regions of abnormally high signal on  $T_2$ -weighted images, and which occurs in many healthy older subjects as well as some stroke patients. Jones *et al.* (1999) demonstrated higher diffusivity and lower anisotropy in areas of leukoaraiosis than in normal tissue, and showed that a map of MD highlights the distinction between leukoaraiosis and the ventricles better than a  $T_2$ -weighted image. More recently, Bastin *et al.* (2007) further demonstrated that FA correlates strongly with magnetisation transfer ratio<sup>d</sup> in regions of leukoaraiosis, but not in comparable normal-appearing white matter, indicating that the loss of white matter integrity in such regions may be tied to a breakdown in myelination.

Due to the demands of **q**-space imaging on MRI hardware, it has been used far less than other forms of dMRI in the clinical domain. Those studies that have employed the technique have been required to essentially abandon the narrow gradient pulse requirement—Assaf *et al.* (2002, 2005) used the parameters  $\delta = 65$  ms and  $\Delta = 71$  ms, at a  $b$ -value equivalent of  $14,000 \text{ s mm}^{-2}$ ; compared with  $353,000 \text{ s mm}^{-2}$  in a true **q**-space experiment (Biton *et al.*, 2006), and just  $1000 \text{ s mm}^{-2}$  in a typical DTI acquisition. However, it has been shown that even under these circumstances, meaningful information about the displacement distribution can be recovered (Lori *et al.*, 2003).

Although dMRI is unique as a technique for studying structural connectivity and white matter integrity, functional magnetic resonance imaging (fMRI), which gives an indication of the level of activity across the brain, provides complementary information. By looking for consistent patterns of correlated activity in different parts of the brain, a degree of *functional* connectivity between regions can be inferred. There have been a number of attempts to combine fMRI and dMRI data acquired from the same subject together (e.g. Cherubini *et al.*, 2007; Guye *et al.*, 2003; Staempfli *et al.*, 2008), and this is likely to remain an active research area for some time.

## 4.6 Summary

We have discussed the physical process of diffusion, and the means by which diffusion displacement distributions of varying complexity can be indirectly measured with MRI. A number of scalar indices indicating the shape of the diffusion tensor have been described—notably the widely used fractional anisotropy. The uses to which these methods have been put in the clinic, including studies of ageing and stroke, have also been briefly surveyed. The existence and measurability of anisotropic diffusion in the white matter of the brain are crucial prerequisites for dMRI-based tractography; and it is to that application that we turn next.

---

<sup>d</sup>The magnetisation transfer ratio is a metric derived from magnetisation transfer MRI, a method which has not been described above. It is sensitive to changes in large molecules such as myelin.





# White matter fibre tracking

THE LOCAL white matter orientation information provided indirectly by dMRI can be used to reconstruct the pathways of major white matter structures through the brain. This reconstruction process is known as fibre tracking, or tractography. A considerable number of tractography algorithms have been put forward, however, which differ in the ways that they interpret the original data, how they handle uncertainty, and how they represent the reconstructed tract. In some cases nontensorial models of diffusion have been employed to handle some of the degeneracies that the diffusion tensor model faces.

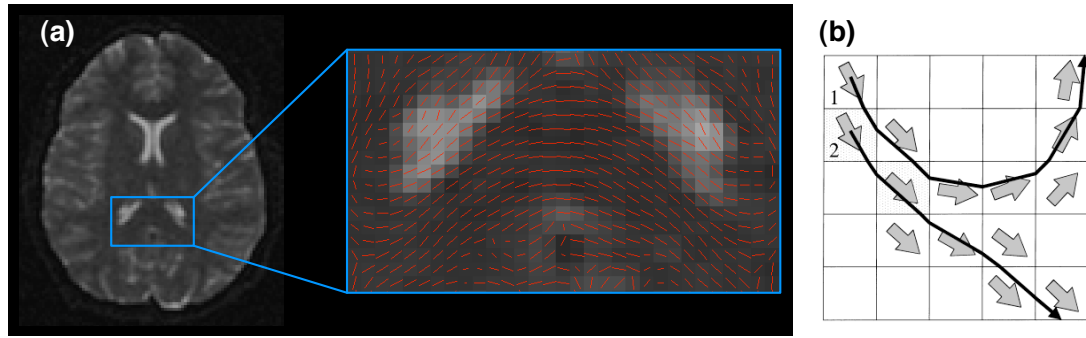
In this chapter we review a number of different types of tractography algorithm, describe their relative advantages and disadvantages, and discuss some of the uses to which fibre tracking methods have been applied. We also mention some of the limitations that still apply to the state of the art algorithms.

## 5.1 Streamlines

We have seen in chapter 4 that the tensor model of diffusion provides an indication of the principal orientation and magnitude of diffusion at a point, in the form of the first eigenvector and associated eigenvalue. This information can be visualised simply by drawing a line, whose orientation and length indicate these two properties, at each location where the model is evaluated—typically a voxel. The components of this representation in a single axial ( $x$ - $y$ ) plane are shown in Fig. 5.1(a). It can be seen by inspection from this figure that there is a fairly smooth curvature in successive principal diffusion direction vectors as they progress across, in this case, the corpus callosum splenium. The most intuitive way to reconstruct a tract is, then, to link these directions together to form a **streamline**. This is the approach taken, in some form, by a majority of tractography algorithms.

Fig. 5.1(b) demonstrates the tract reconstruction process of the Fibre Assignment by Continuous Tracking (FACT) algorithm, which was first demonstrated for fixed rat brain tissue (Mori *et al.*, 1999; Xue *et al.*, 1999). Beginning at the centre of a **seed voxel**, the algorithm moves in the direction of the principal diffusion orientation until reaching the boundary with another voxel, at which time the direction of the reconstructed tract changes to match the orientation of diffusion in the voxel it is entering. This process continues until a termination criterion is met, and is then repeated in the opposite direction from the seed point. It should be noted that the arrowheads shown at each voxel are present for the benefit of interpretation only—they have no physical significance, since diffusion orientation information is directionally nonspecific.

The differences between the early tractography algorithms are primarily in the choice of termination criteria and sampling policy. While FACT samples a trajectory direction exactly once per voxel, other approaches interpolate the original data to obtain local orientation information at a shorter scale, with the reconstruction typically involving short steps of a fixed distance (Basser *et al.*, 2000; Conturo *et al.*, 1999)—a strategy which results in smoother tract pathways than the FACT one. Meanwhile, anisotropy and tract curvature thresholds are commonly used as termination criteria for the reconstruction process, both of which help avoid tracking into



**Figure 5.1:** (a) Visualisation of the principal orientations and magnitudes of diffusion at each voxel in part of a dMRI image. (b) Reconstructing fibre pathways using the FACT algorithm. Subfigure (b) is reproduced from Xue *et al.* (1999).

grey matter or cerebrospinal fluid regions. Further discussion of these issues can be found in a review of fibre tracking methods by Mori & van Zijl (2002).

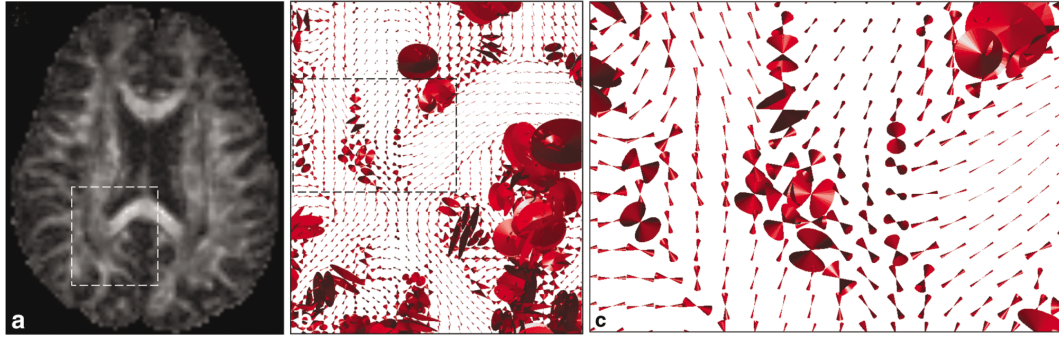
The simplest fibre tracking algorithms are completely deterministic—the principal eigenvector of the diffusion tensor is assumed to be a reliable and noise-free indicator of the local white matter trajectory. The problem, of course, is that the principal diffusion direction is neither of these things. Its reliability is never perfect, and will be affected by the number of gradient directions applied to the sample and any registration errors that occurred during the alignment of the component images, while noise is in fact omnipresent and will tend to “cause a computed trajectory to hop from tract to tract”, as Basser *et al.* (2000) have pointed out. Moreover, noise errors will accumulate as one moves further and further from the seed point. One way to try to circumvent this issue is to impose constraints on the tract reconstruction process which are informed by *a priori* knowledge about the geometry or topology of the underlying fasciculi (Conturo *et al.*, 1999; Poupon *et al.*, 2000). The benefit of these methods—as well as the extent of the problem that they attempt to tackle—is, however, difficult to predict in general terms, since the effects of noise (say) will depend on the particular protocol used to acquire the data, the shape of the tract, the signal-to-noise ratio, the anisotropy characteristics of the tissue, and so on. The final streamline itself gives no indication of the level of confidence that one can expect in the reconstruction.

More recently, streamline-based algorithms have been developed that attempt to indicate the variability that can result, when tracking from a single seed point, due to noise and uncertainty in the data. Some of these techniques are parametric, using a model to explain the data, while others are nonparametric, and therefore implicitly take any source of variability in the results into account. Fundamentally, however, all of these probabilistic approaches are based on the idea of replacing the single principal diffusion direction with a distribution over orientations, which indicates the uncertainty associated with the data at each voxel. One can then generate a family of streamlines from a single seed point using a Monte Carlo approach, sampling from these local distributions each time a new tracking direction is needed. Early work in this vein was published by Lazar & Alexander (2002) and Parker *et al.* (2003), who used the tensor shape to choose the variance of the orientation distributions. The approach demonstrated by Parker *et al.* was later developed further by Cook *et al.* (2004).

Bootstrap approaches to tractography are an example of a nonparametric statistical approach. Bootstrap is a resampling method, which requires that multiple measurements of the diffusion-weighted signal be taken for each diffusion gradient direction. Then, rather than using all of these data to fit a single diffusion tensor—which is the maximum likelihood approach taken by more simplistic algorithms—a subset of the data is sampled, with replacement, from the multiple measurements, and the tensor is calculated from this subset. A large number of these subsets are then extracted from this original data set, producing an empirical distribution over each of the free parameters in the diffusion tensor model. A general approach to using bootstrap to characterise uncertainty in dMRI data was put forward by Pajevic & Basser (2003),

Full data set				Sample 1	Sample 2
$A_1^{(1)}$	$A_1^{(2)}$	$\dots$	$A_1^{(6)}$	$A_1^{(2)}$	$A_1^{(4)}$
$A_2^{(1)}$	$A_2^{(2)}$	$\dots$	$A_2^{(6)}$	$A_2^{(5)}$	$A_2^{(4)}$
$A_3^{(1)}$	$A_3^{(2)}$	$\dots$	$A_3^{(6)}$	$A_3^{(1)}$	$A_3^{(1)}$
$A_4^{(1)}$	$A_4^{(2)}$	$\dots$	$A_4^{(6)}$	$A_4^{(2)}$	$A_4^{(6)}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
$\underbrace{\hspace{10em}}$ regression $\downarrow$ $\hat{D}$				$\underbrace{\hspace{10em}}$ $\downarrow$ $D^{(1)}$	$\underbrace{\hspace{10em}}$ $\downarrow$ $D^{(2)}$

**Table 5.1:** Illustration of the application of bootstrapping to a dMRI data set containing repeated measurements. We denote the  $i$ th signal measurement using the  $k$ th gradient direction as  $A_k^{(i)}$ .

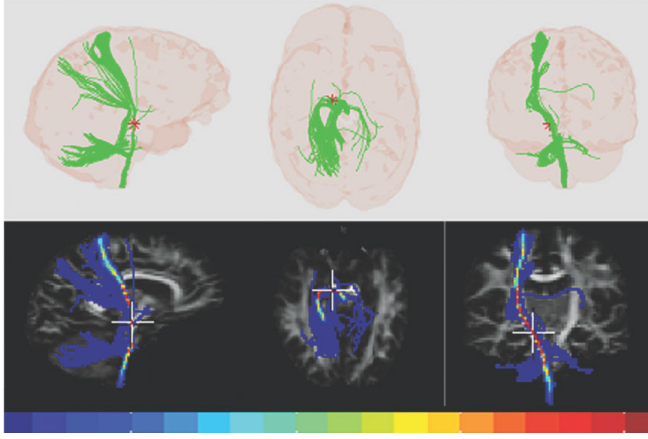


**Figure 5.2:** Orientation uncertainty in dMRI data, visualised as cones showing the 95% angular confidence interval at each voxel. Subfigure (b) corresponds to the area of (a) indicated with a box; likewise the further enlarged image (c). Reproduced from Jones (2003).

with applications to tractography following later (Jones & Pierpaoli, 2005; Lazar & Alexander, 2005).

Let us assume, for the sake of argument, that we have made six signal measurements for each of the gradient directions applied during a dMRI experiment. The diffusion tensor,  $D$ , can then be estimated from various subsets of these data, provided that at least the minimum six noncollinear gradient directions, plus a measurement with no diffusion weighting, contribute data to each subset. This is the principle employed by Jones (2003), and illustrated by Table 5.1. The maximum likelihood tensor is denoted by  $\hat{D}$ , while those estimated by sampled subsets of the data are denoted  $D^{(1)}$  and so on. The latter can be used to estimate the uncertainty associated with the principal eigenvector, which is visualised in Fig. 5.2. Each set of sampled tensors for a given brain volume can then be used, in turn, to generate a single streamline from a chosen seed point, using a normal deterministic algorithm. The result will be a set of streamlines with a spatial distribution that reflects the variability encountered by the streamlining algorithm across the sample set—as shown in Fig. 5.3.

There are some interesting characteristics of the uncertainty elucidated in this way. Firstly, we can see by immediate inspection of Fig. 5.2 that the width of the 95% confidence interval on the principal diffusion orientation, which is depicted there, is highly variable between voxels. Near the middle of the corpus callosum splenium the confidence interval is extremely narrow. In this region, the maximum likelihood tensor would provide a reliable indication of the trajectory of this white matter structure. By contrast, the uncertainty is huge in areas which are composed primarily of *csf*—like near the bottom right of subfigure (b)—where



**Figure 5.3:** Results of applying the bootstrap method to tractography of the corticospinal tract. From a single seed point, which is indicated on each subfigure, a number of sample streamlines are produced (top). The proportion of the streamlines visiting each voxel can be counted to form a “visitation map” (bottom). Reproduced from Jones & Pierpaoli (2005).

diffusion is close to isotropic. Less predictable, however, is the effect of fibre crossings, which can be observed near the centre of (c). In this case, diffusion is approximately oblate, with two relatively large eigenvectors and one smaller one; and so the principal direction is less certain. The cone metaphor reflects this.

The necessity of acquiring multiple signal measurements for each diffusion gradient direction represents a problem for the basic bootstrap paradigm, because it will result in considerably extended scanning times without the improvement in angular resolution that would result from spending this time sampling more directions. Long scan times are particularly problematic in the clinical domain, since patients cannot be expected to remain still for long periods of time. Furthermore, the bootstrap method can substantially underestimate the degree of uncertainty in the tensor components when the number of repeated acquisitions is small (Chung *et al.*, 2006). However, a method known as the wild bootstrap offers to remove the need for multiple acquisitions when estimating the uncertainty in dMRI data (Whitcher *et al.*, 2007).

The wild bootstrap differs from “ordinary” bootstrapping in that it works with the residuals from a diffusion tensor fit to the signal data. If we describe a vector of unknown parameters,  $\mathbf{x} = (D_{xx}, D_{yy}, D_{zz}, D_{xy}, D_{xz}, D_{yz}, \ln A_0)$ —where  $D_{xx}$  (and so on) are the tensor components, and  $A_0$  is the signal without diffusion weighting—then the linear model used to estimate these components can be written out as

$$\mathbf{A} = \mathbf{B}\mathbf{x} + \boldsymbol{\epsilon},$$

where  $\mathbf{A}$  is a vector of observed log-signal values,  $\mathbf{B}$  is a matrix describing the diffusion gradient directions applied, and  $\boldsymbol{\epsilon}$  is a vector of error terms. Thus we can evaluate an estimate for the parameters,  $\hat{\mathbf{x}}$ , using least-squares regression and our knowledge of  $\mathbf{A}$  and  $\mathbf{B}$ . As with other bootstrap methods, we do not need an explicit model for the errors, which are caused by noise and misregistration and so on. However, we subsequently use them to generate samples according to

$$A_k^{(i)} = \mathbf{B}_k \cdot \hat{\mathbf{x}} + h_k s_k^{(i)} \epsilon_k, \quad (5.1)$$

where  $\mathbf{B}_k$  and  $\epsilon_k$  are the elements of  $\mathbf{B}$  and  $\boldsymbol{\epsilon}$  corresponding to the  $k$ th direction, and  $s_k^{(i)}$  has the simple probability mass function

$$\Pr(s_k^{(i)} = s) = \begin{cases} \frac{1}{2} & \text{for } s = \pm 1 \\ 0 & \text{otherwise} \end{cases} \quad \forall k, i. \quad (5.2)$$

The constant  $h_k$  in Eq. (5.1) is used to ensure that the sampled residuals have the covariance structure required by the method (see Chung *et al.*, 2006, for details). Rather than repeatedly measure  $A_k$ , therefore, we instead resample the data by randomly permuting the signs of the residuals—i.e. by sampling from Eq. (5.2) for each value of  $i$  and  $k$ . Thus, only a single set of real measurements need be made, keeping scanning time short.

## 5.2 BEDPOST

It should be noted that the wild bootstrap introduces a dependence on the diffusion tensor model which is not present using ordinary bootstrap. Since the acquired data must be fitted to some kind of model for residuals to be available, the wild bootstrap is by nature a model-based resampling method. However, while the signal measured for each diffusion gradient applied is modelled using the diffusion tensor formalism, no model is used to explain the variability itself.

It is possible to go further, and model the observed data including their inherent uncertainty. This is the aim of another category of tractography algorithms, including the BEDPOST algorithm (Bayesian Estimation of Diffusion Parameters Obtained using Sampling Techniques; see Behrens *et al.*, 2003b), which has been used for most of the practical parts of this thesis. By way of illustration of a fully model-based approach to tractography, and because of its centrality to work described later, this algorithm is fully described below.

The BEDPOST algorithm uses Markov chain Monte Carlo sampling to estimate diffusion MRI parameters. As above, the algorithm works with a vector of observed log-signal data,  $\mathbf{A}$ , and a model parameter vector,  $\mathbf{x}$ . However, because the diffusion tensor model can only usefully describe a single principal diffusion direction—since the second eigenvector is constrained to be orthogonal to the first—similar information can be embodied in a simpler model. In particular, Behrens *et al.* assume that the diffusion displacement distribution is a mixture of two Gaussians, in which one “compartment” is isotropic and the other is perfectly anisotropic, describing a single local tract orientation. The signal for the  $k$ th diffusion direction,  $\mu_k$ , is then given by

$$\mu_k = A_0 \left( (1 - f) \exp(-b_k D) + f \exp(-b_k D \mathbf{G}_k^T \mathbf{R} \mathbf{M} \mathbf{R}^T \mathbf{G}_k) \right), \quad (5.3)$$

where  $b_k$  is the  $k$ th scalar  $b$ -value,  $\mathbf{G}_k$  is the  $k$ th diffusion encoding direction represented as a column vector,

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

and  $\mathbf{R}$  rotates  $\mathbf{M}$  to align with the fibre direction in the voxel, which requires two implicit angles  $(\theta, \phi)$ . Compare Eq. (5.3) with the standard dMRI formulations in Eqs (4.7) and (4.8). The natural index of anisotropy arising from this model is the mixture coefficient,  $f$ , which we refer to as the **anisotropic volume fraction** (AVF). Note that the model provides no information about anisotropy perpendicular to the direction encoded by  $\mathbf{R}$ ; but then such information is not directly relevant to streamline tractography.

Under the generative model of local diffusion described by Eq. (5.3) and the assumption that noise is independent and identically distributed for each measurement, the likelihood of the observed data is given by

$$P(\mathbf{A}|\mathbf{x}) = \prod_k P(A_k|\mathbf{x}), \quad (5.4)$$

where

$$P(A_k|\mathbf{x}) \sim N(\mu_k, \sigma^2); \quad (5.5)$$

and so the full parameter vector is  $\mathbf{x} = (A_0, D, f, \theta, \phi, \sigma)$ . The posterior distribution over these parameters is given by Bayes’ rule:

$$P(\mathbf{x}|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{x})P(\mathbf{x})}{\int P(\mathbf{A}|\mathbf{x}')P(\mathbf{x}')d\mathbf{x}'} . \quad (5.6)$$

For the purposes of fibre tracking, however, the most important parameters at each voxel are the angles which provide tract orientation information. If we wish to obtain distributions over  $\mathbf{x}_1 = (\theta, \phi)$ , we will need to calculate the marginal distribution given by

$$P(\mathbf{x}_1|\mathbf{A}) = \int P(\mathbf{x}|\mathbf{A})d\mathbf{x}_2, \quad (5.7)$$

where  $\mathbf{x}_2 = (A_0, D, f, \sigma)$ , a vector consisting of the remaining parameters. Both the evidence term in the denominator of Eq. (5.6) and the marginal distribution of Eq. (5.7) require the evaluation of complex integrals, however, and cannot be expected to be soluble analytically. We therefore turn to MCMC sampling to evaluate them empirically.

The priors,  $P(\mathbf{x})$ , in Eq. (5.6) are chosen by the authors to be uninformative, except where ensuring positivity is appropriate: in  $A_0$  and  $f$ . Initialisation for the Markov chains is provided by performing a normal least-squares diffusion tensor fit to the data at each voxel, and using tensor analogues of each parameter. Samples for  $\sigma$  are generated using a Gibbs sampler, and all other parameters are sampled using the Metropolis–Hastings algorithm. Proposal distributions for the latter are zero-mean Gaussians whose variance is tuned to maintain an acceptance rate of 0.5.

The generative model for the noisy data, Eq. (5.5), takes the form of a normal distribution with known mean—given knowledge of the partial parameter vector  $\mathbf{x}_3 = (A_0, D, f, \theta, \phi)$ —and unknown variance. This is a common and therefore well-characterised situation. Using a gamma prior distribution for the precision,  $\tau = 1/\sigma^2$ , viz.

$$P(\tau|\alpha, \beta) = \text{Gamma}(\alpha, \beta) = \frac{\tau^{\alpha-1} \beta^\alpha e^{-\beta\tau}}{\Gamma(\alpha)},$$

where  $\Gamma(\cdot)$  is the gamma function, the posterior over  $\tau$  given data  $\mathbf{A}$  is another gamma distribution:

$$P(\tau|\alpha, \beta, \mathbf{A}, \mathbf{x}_3) = \text{Gamma}\left(\alpha + \frac{K}{2}, \beta + \frac{1}{2} \sum_k (A_k - \mu_k)^2\right),$$

where  $K$  is the total number of gradient directions acquired. This is used by BEDPOST as the conditional distribution for the Gibbs sampler, although the authors do not explicitly state how they chose the prior hyperparameters  $\alpha$  and  $\beta$ .

The marginal distribution for  $\mathbf{x}_1$  is trivially extracted from the samples over  $\mathbf{x}$  by considering only  $\theta$  and  $\phi$  from each sample vector. The tractography part of the algorithm—which the authors call ProbTrack—then uses these samples to reconstruct a set of “probabilistic streamlines” using a normal streamlining approach. Given a seed voxel,  $\mathbf{a}$ , the process is as follows.

1. Start with the current “front” of the streamline set to  $\mathbf{a}$ .
2. Select a random sample,  $(\theta, \phi)$ , from  $P(\theta, \phi|\mathbf{A})$  at the streamline front.
3. Move the front some small distance in the direction of  $(\theta, \phi)$ .
4. Return to step 2, and repeat until a stopping criterion is met.

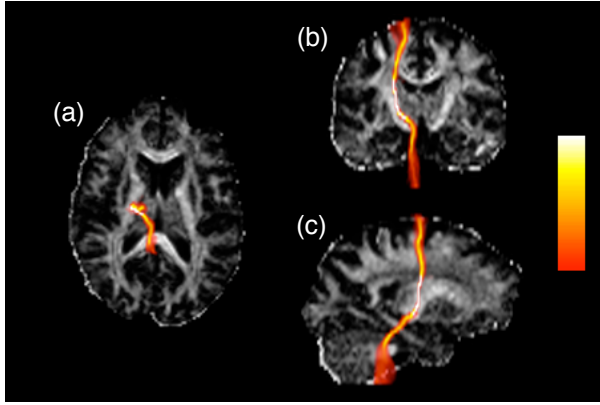
The stopping criteria are not strict, stipulating only that a streamline is not allowed to curve by more than about  $80^\circ$ , and that a streamline will be terminated if it leaves the brain or enters an area that it has already visited.

To evaluate the direction of propagation at any location in the brain, not just those that coincide with voxel centre points, some kind of interpolation scheme is required. The authors use a probabilistic analogue of trilinear interpolation, in which a sample is drawn from one of the two adjacent voxels in each dimension according to how close the sample location is to each of them. Indexing in voxel steps, the sample location is taken from the p.m.f.

$$\Pr(x = v) = \begin{cases} \text{ceil}(x) - x & \text{for } v = \text{floor}(x) \\ x - \text{floor}(x) & \text{for } v = \text{ceil}(x), \end{cases}$$

where floor and ceil are the usual floor and ceiling functions. If  $x = \text{floor}(x)$ —that is,  $x$  falls exactly on a voxel location—then the sample is taken from that voxel with unit probability.

This procedure for generating streamlines is repeated a large number of times (typically 5000) for a particular seed point, generating a spatial distribution for the tract running through the seed point at  $\mathbf{a}$ . This distribution may be usefully discretised by counting up the number of streamlines passing through each voxel and associating this count with the voxel volume. An



**Figure 5.4:** Example discretised spatial distribution from the BEDPOST/ProbTrack tractography algorithm, showing the corticospinal tract in axial (a), coronal (b) and sagittal (c) maximum intensity projections. The underlying greyscale image shows AVF in the slice in-plane with the seed point in each case. White indicates that nearly all streamlines pass through the local voxel, while red means that very few do. The full colour scale is shown.

example of the result is shown in Fig. 5.4. These data can be interpreted as confidence bounds on the location of the most probable tract passing through the seed point.

Behrens *et al.* showed, in their paper, that the levels of uncertainty estimated by their method are comparable with those estimated by the bootstrap approach described by Jones (2003)—thus justifying, to some extent, the additional assumptions that they make in their fully model-based approach. The advantage of this added model specificity, meanwhile, is an improved sensitivity.

A standard implementation of BEDPOST is freely available as part of the FSL package of software tools (Smith *et al.*, 2004), which is written and maintained by the FMRIB centre at the University of Oxford.

A number of variations on, and extensions of, the BEDPOST method have been proposed. Friman *et al.* (2006) describe another alternative model for diffusion at a voxel, which is essentially the tensor model, but with the two smaller eigenvalues constrained to be equal—that is,  $\lambda_2 = \lambda_3 = \alpha$ —thereby producing the form

$$\mu_k = A_0 \exp(-\alpha b_k) \exp\left(-(\lambda_1 - \alpha) b_k \mathbf{G}_k^T \mathbf{R} \mathbf{M} \mathbf{R}^T \mathbf{G}_k\right). \quad (5.8)$$

The authors also use a more theoretically justified noise model, whose variance depends on the signal value; and they use point estimates for the “nuisance” parameters in the model in order to reduce its computational demands.

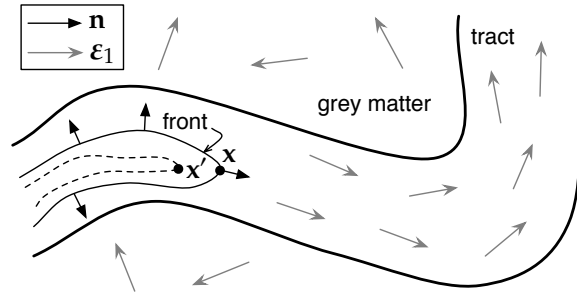
Neither the compartment model described by Eq. (5.3) nor the constrained model of Eq. (5.8) can account for more than one fibre orientation at a voxel. Rather, multiple fibre orientation information is manifested as increased uncertainty in the single orientation that they can represent. However, both models can be generalised to handle this case, which occurs commonly in the brain at typical imaging resolution—the compartment model by adding extra anisotropic compartments (Behrens *et al.*, 2007), and the constrained case by modelling additional tensors (Hosey *et al.*, 2005). It is generally wise to use the simplest model that explains the data satisfactorily at each voxel, rather than simply to fit multiple fibre orientations at every location in the brain. Hosey *et al.* achieve this by fitting one and two tensor models at each voxel and using probabilistic model selection to choose between the results, while Behrens *et al.* fit a single, complex model but apply a technique known as automatic relevance determination (see MacKay, 1995, §7) to factor out unneeded parameters.

### 5.3 Fast marching

Streamline generation is not the basis for all fibre tracking algorithms; although it is, as we have mentioned, the most common. One alternative general approach is to propagate a 3-D surface or front in all directions from the seed point at once, such that its speed is faster in some directions than in others—a method called *fast marching tractography* (FMT; see Parker *et al.*, 2002b). A speed function is used to define how fast the front moves as it progresses through



**Figure 5.5:** Front propagation in fast marching tractography. The speed function,  $F$ , is designed so that the front of the spreading region grows most quickly where its normal vector,  $\mathbf{n}$ , aligns closely with the principal eigenvector of the local diffusion tensor,  $\epsilon_1$ . Thus the front will move fastest along paths with smoothly varying principal diffusion orientation. Arrowheads on the eigenvectors are notional. After Parker *et al.* (2002b).



the brain. Parker *et al.* (2002a) use the speed function

$$F(\mathbf{x}) = \min \{ |\epsilon_1(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})|, |\epsilon_1(\mathbf{x}') \cdot \mathbf{n}(\mathbf{x})|, |\epsilon_1(\mathbf{x}) \cdot \epsilon_1(\mathbf{x}')| \}, \quad (5.9)$$

where  $\mathbf{n}(\mathbf{x})$  is the local normal to the front at point  $\mathbf{x}$ , and  $\epsilon_1(\mathbf{x})$  is the first eigenvector of the local diffusion tensor. The point  $\mathbf{x}'$  represents the position of a neighbouring voxel that has already been passed by the front. These terms are visualised in Fig. 5.5.

As we follow this propagating front out from the seed point, we can establish a “time of arrival” for each voxel in the brain. Wherever the front moves fastest, the time of arrival to voxels along its route will be low. One can do target-based tractography by then performing a gradient descent in a time of arrival map, from the target voxel back to the seed. Exploratory tractography from a seed point is also possible by using every other voxel in the brain as a target point in turn, and retaining those pathways which are most plausible under some criterion, such as the minimum or average value of the speed function along them.

The performance of FMT hinges on the choice of speed function. Parker *et al.* (2002b) discuss alternative forms for the speed function, although they limit themselves to the case where the first tensor eigenvector can be considered a reliable indicator of tract direction. Since then, however, Staempfli *et al.* (2006) have described a set of four speed functions, from among which their FMT algorithm selects, depending on the tensor shape at  $\mathbf{x}$  and  $\mathbf{x}'$ . This allows their method to track through regions in which diffusion has an oblate, rather than prolate, profile.

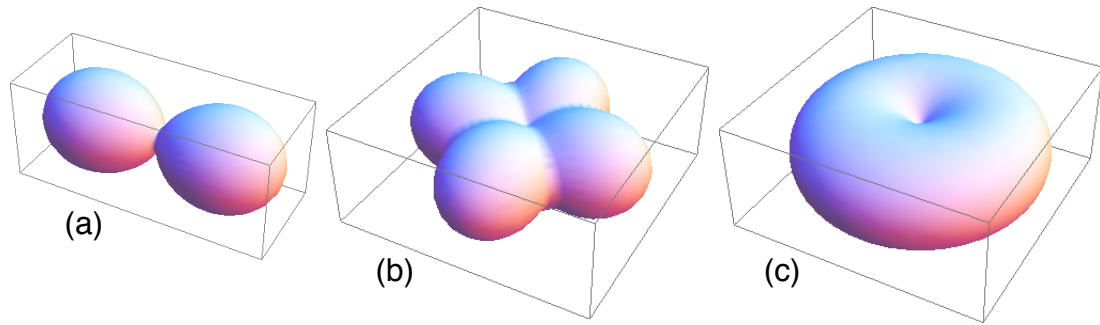
## 5.4 High angular resolution methods

A number of models of diffusion have been developed for the purpose of elucidating the orientations of multiple fibre populations within a voxel. Some of these are direct extensions of simpler models, as we have already seen, while others were designed from the outset to work with crossing fibres.

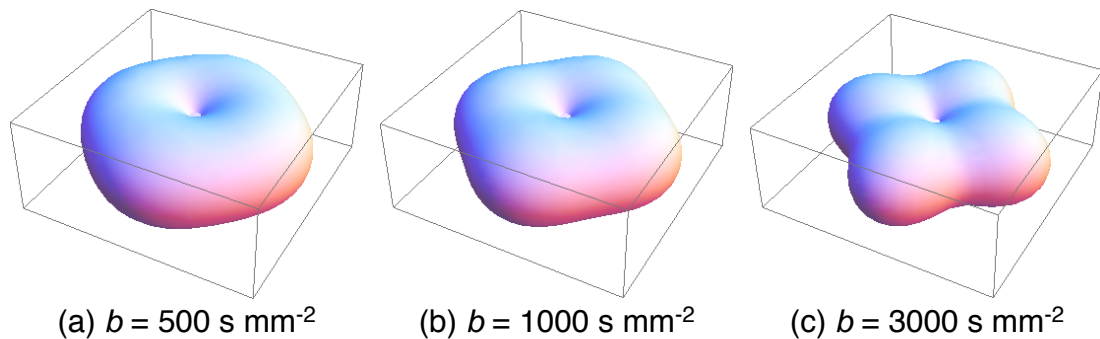
The need for more complex models than the tensor model in tractography has been touched upon earlier in this chapter, but Fig. 5.6 demonstrates the issue explicitly (see also Frank, 2001). With a single fibre orientation per voxel, the tensor model is an adequate model, effectively representing the diffusion profile expected for this case, as in subfigure (a). On the other hand, we would like to be able to recover a profile encapsulating two fibre orientations when this is justified (b), but instead the tensor can only represent a directionally nonspecific profile (c). In order to track effectively through regions of crossing fibres, however, the structure in the inherent diffusion profile must be retained.

The first requirement for successful elucidation of crossing fibre architectures is, then, a diffusion model that is capable of representing their relatively complex structure; but there are also commensurate acquisition requirements. Since more complex models have more parameters, and in particular because they aim to more fully represent the diffusion profile, larger numbers of gradients must be applied to improve the angular resolution of the scan. For this reason, the modelling and acquisition techniques that aim to represent complex intravoxel architectures are called high angular resolution diffusion imaging (HARDI) methods. Secondly, in order to produce strong enough contrast between the signal effects of each fibre population, greater diffusion weighting—corresponding to greater values of  $b$ , the weighting coefficient—is usually applied. This can be achieved by increasing gradient strength or diffusion time.





**Figure 5.6:** Visualisation of diffusivity as a function of gradient orientation. When there is a single fibre population within a voxel, it produces a diffusivity profile like the one in (a), which is well represented by the diffusion tensor model. In the presence of two orthogonal populations, the true profile is something like (b), but a single diffusion tensor is only capable of representing the ambiguous case shown in (c).



**Figure 5.7:** Dependence of the signal on  $b$ -value. Multiple fibre orientations are better contrasted at higher levels of diffusion weighting. Peak diffusivity in each of the two component tensors was  $7.5 \times 10^{-4} \text{ mm}^2 \text{ s}^{-1}$ .

The effect on angular contrast of increasing the  $b$ -value is shown in Fig. 5.7, and described in Alexander *et al.* (2001). It should be borne in mind that unfortunately, higher  $b$ -values also produce less overall signal—since the weighting factor determines the level of attenuation in the signal due to diffusion effects—so the signal-to-noise ratio of the acquisition is lower.

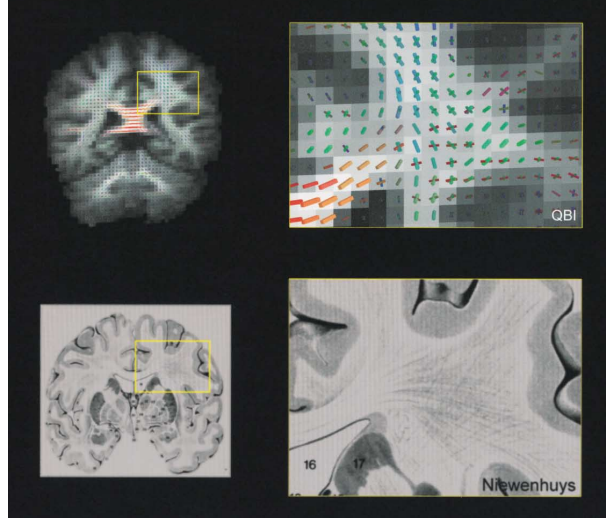
One way to handle multiple fibre directions is to use multiple tensors (Tuch *et al.*, 2002)—an approach we have already seen employed, in a constrained form, by Hosey *et al.* (2005). Under this model, the diffusion displacement distribution is assumed to be a mixture of Gaussians with different covariance structures. Two tensors are able to faithfully represent the situation shown in Fig. 5.6(b), although a third tensor would need to be used for the case of three fibre populations, and it is often not possible to know *a priori* how many fibre populations are expected within a given voxel.

It should be noted that a multiple tensor model assumes that the diffusing water molecules do not move between fibre populations during the course of the experiment. This is known as the assumption of *slow exchange*, and it is a typical assumption in the analysis of crossing fibre structure. Since the root-mean-squared diffusion distance for a typical dMRI protocol is of the order of  $10 \text{ } \mu\text{m}$ , compared to a typical axon diameter of a few microns, this assumption is thought to be a reasonable one for most purposes. Fibre tracts consist of bundles of hundreds of axons, and so diffusion over the width of a few axons will rarely exchange between bundles.

### 5.4.1 Using $q$ -space

An alternative general approach to the crossing fibre problem is to employ  $q$ -space diffusion imaging. As we saw in §4.3,  $q$ -space imaging allows us to recover an arbitrary displacement

**Figure 5.8:** Results from  $\mathbf{q}$ -ball imaging and comparison with invasive tracing. The  $\mathbf{q}$ -ball reconstruction of fibre orientations effectively represents the fanning out of pathways emerging from part of the corpus callosum, which can be seen in an atlas of the central nervous system (Nieuwenhuys, 1996). Reproduced from Tuch *et al.* (2003).



distribution in a model-free manner, by taking a Fourier transform of MRI signal information acquired using an appropriate scheme. A scheme suitable for recovering crossing fibre orientations was described by Wedeen *et al.* (2005), using 515  $\mathbf{q}$ -vectors and a maximal  $b$ -value equivalent of  $17,000 \text{ s mm}^{-2}$ . Having recovered a spatial displacement distribution,  $P(\mathbf{r})$ , an *orientation distribution function* (ODF) can be calculated by projecting the distribution onto the unit sphere. That is,

$$\Psi(\hat{\mathbf{r}}) = \int_0^\infty P(\rho\hat{\mathbf{r}})\rho^2 d\rho, \quad (5.10)$$

where  $\mathbf{r} = \rho\hat{\mathbf{r}}$ . In this case the authors use the squared vector length,  $\rho^2$ , as a weighting factor. The ODF then provides the information needed to perform tractography, using a streamline method or otherwise, in the region. (It should be noted, however, that the ODF has no probabilistic interpretation because it is not properly normalised.) This approach is called *diffusion spectrum imaging* (DSI).

The biggest problem with DSI is its acquisition requirements. The protocol makes no real attempt to satisfy the narrow gradient pulse assumption, so the demands it makes on gradient hardware are not extreme; but because it samples  $\mathbf{q}$ -space quite thoroughly, imaging a brain volume at a reasonable resolution takes far longer than a comparable DTI protocol.

A step towards reduction of the  $\mathbf{q}$ -space sampling requirements of DSI was taken by the development of so-called  $\mathbf{q}$ -ball imaging (Tuch *et al.*, 2003; Tuch, 2004). In this case, the length of the sampled  $\mathbf{q}$ -vectors is fixed so that they lie on a sphere. The authors show that an ODF can then be recovered directly by means of an integral transform called the Funk–Radon transform, which has its roots in computed tomography, a medical imaging technique using x-rays. The authors also describe a method for calculating this transform that is reasonably simple and computationally inexpensive. Fast marching tractography has since been demonstrated using the  $\mathbf{q}$ -ball ODF as a speed function (Campbell *et al.*, 2005).

It has been shown that the  $\mathbf{q}$ -ball method produces ODF information that is in fairly good agreement with standard, invasive tracing work (see Fig. 5.8), and certainly provides more useful information for tractography in crossing fibre regions than the tensor model (Fig. 5.9). Note that in Fig. 5.9(a) the first tensor eigenvector represents a more or less arbitrary orientation in regions where the crossing occurs. This is consistent with the degenerate representation expected under this model (cf. Fig. 5.6).

Jansons & Alexander (2003) describe an alternative to an orientation distribution function called *persistent angular structure* (PAS). As with the ODF formulation, the aim is to capture the orientation information in the signal which is important for tractography, whilst discarding the less salient radial information. The radial part of the diffusion displacement distribution is

therefore factored out and represented by a Dirac  $\delta$ -function, viz.

$$P(\mathbf{r}) = \frac{p(\hat{\mathbf{r}}) \delta(|\mathbf{r}| - \rho)}{\rho^2},$$

where the function  $p(\hat{\mathbf{r}})$  is the  $\text{PAS}$ , the angular component of the distribution. Here,  $\rho$  is a parameter that has to be chosen independently. By means of an optimisation which is constrained by the relationship between the data and the displacement distribution—embodied in Eq. (4.13)—the authors arrive at the maximum entropy solution

$$p(\hat{\mathbf{r}}) = \exp \left( \lambda_0 + \sum_j \lambda_j \exp(i\mathbf{q}_j \cdot \rho \hat{\mathbf{r}}) \right), \quad (5.11)$$

where  $\{\lambda_j\}$  are constants to be found. The maximum entropy solution is the most uninformative function possible, subject to the constraints imposed by the data. The intuition of this approach is to encode just the angular structure “reported” by the acquired data, without introducing extra information by making additional assumptions.

A big—perhaps the biggest—advantage of  $\text{PAS-MRI}$  is its modest acquisition requirements. Jansons & Alexander use a scheme involving just 54 nonzero  $\mathbf{q}$ -vectors, compared to hundreds for a typical  $\text{DSI}$  or  $\mathbf{q}$ -ball experiment. The trade-off, however, comes in computation time. Since the  $\text{PAS}$ , Eq. (5.11), is a nonlinear combination of functions, reconstruction times for  $\text{PAS-MRI}$  are typically orders of magnitude longer than those required by the other, linear techniques. With present computing power, the time needed to fully process a large data set could be prohibitive.

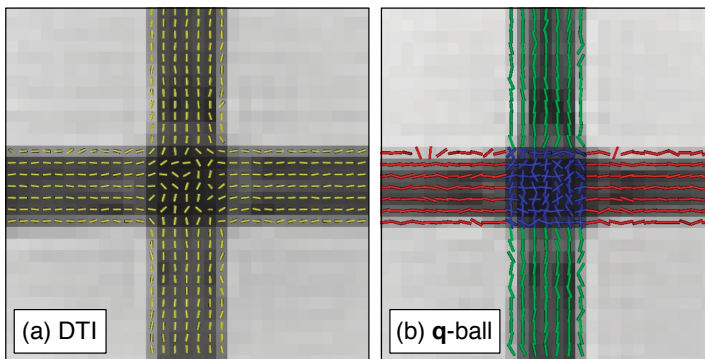
### 5.4.2 Spherical deconvolution

A further subcategory of  $\text{HARDI}$  methods use a technique called spherical deconvolution, which allows one to recover an  $\text{ODF}$  without relying on the Fourier relationship between the  $\text{dMRI}$  signal and the displacement distribution, which is anyway only approximate since the narrow gradient assumption is not fulfilled. Instead, the fundamental assumption here is that the signal arises from the convolution of an  $\text{ODF}$  with a “response function”, which is assumed to be invariant across all white matter in the brain, with partial volume effects accounting for all nonorientational variability (Tournier *et al.*, 2004). Slow exchange is also assumed. We therefore write, for a particular  $b$ -value,

$$A(\theta, \phi) = \sum_i f_i \mathbf{R}_i S(\theta) = \Psi(\theta, \phi) \otimes S(\theta), \quad (5.12)$$

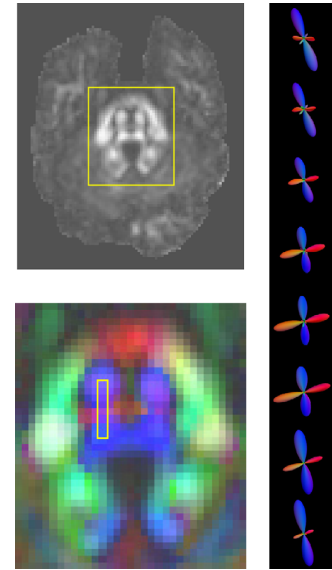
where  $\theta$  represents the polar angle and  $\phi$  the azimuthal angle in spherical polar coordinates,  $f_i$  is the volume fraction of the  $i$ th fibre population, and  $\mathbf{R}_i$  is a rotation matrix representing its orientation. The symbol  $\otimes$  represents convolution on the unit sphere. We note that the unit vector  $\hat{\mathbf{r}}$  used above as the  $\text{ODF}$  parameter is related to the two angles by

$$\hat{\mathbf{r}} = (\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta).$$



**Figure 5.9:** Fibre orientation information reconstructed using DTI and  $\mathbf{q}$ -ball methods for a specially constructed phantom, mimicking orthogonal crossing fibre populations. Adapted from Perrin *et al.* (2005).

**Figure 5.10:** Results of applying the spherical deconvolution method in a region of fibre crossing in the pons. The two fibre orientations appear quite distinct from one another and are qualitatively accurate representations of the underlying architecture. Reproduced from Tournier *et al.* (2004).



The response function,  $S$ , is a function only of  $\theta$  because it is taken to be axially symmetric. Given the knowledge of an appropriate response function, the ODF can therefore be deconvolved out of the signal profile, at least in principle.

It is worth noting that the spherical deconvolution model is related to the anisotropic component of the BEDPOST partial volume model, Eq. (5.3). In that case the response function represents Gaussian diffusion along a single orientation, and the ODF is a  $\delta$ -function whose orientation corresponds to that of the modelled fibre pathway. PAS-MRI can also be framed as a deconvolution (Alexander, 2005).

Tournier *et al.*, by contrast, use an ODF which can represent multiple directions; and they represent it, along with the response function, in terms of a set of functions known as the spherical harmonics (Riley *et al.*, 2002). These functions form an orthonormal basis set over the sphere, and their use in general spherical deconvolution problems has been described by Healy *et al.* (1998). Representation of the signal profile,  $A$ , using these basis functions had already been described (Alexander *et al.*, 2002; Frank, 2002). Under this parameterisation, the ODF can be recovered by means of a straightforward set of matrix multiplications, given knowledge of the response function—which the authors establish by observing the signal profile in strongly anisotropic parts of the brain.

The method has been demonstrated to work well for resolving fibre crossings in simulations and in real data acquired with a modest 60 gradient directions at a  $b$ -value of about  $3000 \text{ s mm}^{-2}$  (see Fig. 5.10). The authors estimate that using these acquisition parameters, two fibre orientations with a separation of  $60^\circ$  can be recovered with a standard deviation of around  $9^\circ$ . The minimum resolvable separation is estimated to be about  $40^\circ$ .

The validity of the assumption of equivalent response throughout the brain is hard to establish, but the most significant shortcoming of the method is probably its sensitivity to artefacts caused by noise. Recently developed methods for regularising the ODF (Sakaie & Lowe, 2007; Tournier *et al.*, 2007) promise to mitigate this issue significantly, however—even when the signal-to-noise ratio is low. Thus it may be possible to apply the method to recover useful orientation information even at the lower  $b$ -values commonly used in DTI experiments. A parametric version of the spherical deconvolution method has also been recently developed, allowing Bayesian statistics to be used to infer an ODF (Kaden *et al.*, 2007).

## 5.5 Applications and challenges

Given the increasingly formidable array of ideas and innovations which have been thrown at the fibre tracking problem, it is natural to ask what scientific uses there may be for reliable

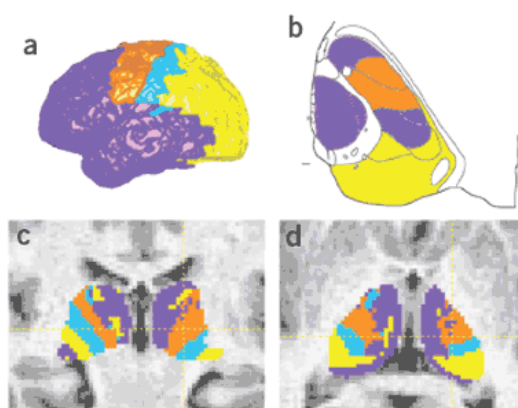
tractography methods once they have been developed. At present, there are two general categories of application for these algorithms which have appeared in the literature.

The first application might be loosely described as connectivity analysis. Despite the fact that tractography is still very much a field in its infancy, it is already beginning to provide information about the brain's internal connections which are corroborating the findings of more well established—and more invasive—neuroscientific techniques. In an impressive piece of work, Behrens *et al.* (2003a) demonstrated, using tractography, that voxels in the thalamus can be effectively categorised by the targets of their most likely projections into cortex. The resulting thalamic parcellations are in close agreement with atlas data (see Fig. 5.11), and have been further reinforced by functional MRI results (Johansen-Berg *et al.*, 2005). Similar principles have been applied to the corpus callosum (Huang *et al.*, 2005), and used to identify boundaries between cortical regions based on their connectivity (Johansen-Berg *et al.*, 2004).

The second category of application encompasses the segmentation and visualisation of specific tracts. The emphasis in this case is more clinical than neuroscientific, since segmenting a particular tract is often a precursor to comparative analysis of anisotropy—or some other indicator of pathology—between a patient group and controls. We will not expand further on the segmentation application here, however, because it will be the focus of the next three chapters; and therefore will be described fully elsewhere. Tract visualisation can be useful in its own right as a preoperative surgical planning tool, since any invasive treatment will naturally try to minimise damage to important connective pathways—although at present it is highly advisable to avoid setting too much store by tractographic results in such critical applications (Kinoshita *et al.*, 2005).

Notwithstanding their increasing popularity and promising early results, tractography methods have some outstanding theoretical and practical limitations. The problem of handling crossing fibres cannot be said to be fully solved, especially in the relatively high noise and low angular resolution regime which is common in clinical MRI scanning. There is also an additional degeneracy which is widely recognised, but whose impact has not yet been fully characterised: the problem of “kissing” fibres (Basser *et al.*, 2000). From a fibre tracking point of view, it is important to be able to distinguish the two intravoxel architectures shown in Fig. 5.12, but a recovered ODF will usually not provide enough information to do so.

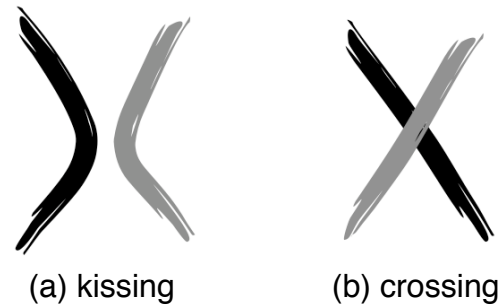
However plausible the reconstructed tracts may appear to be, the issue of validation is a significant one. Efforts to validate tractography methods have recently increased, and include computer simulation work (as in Hosey *et al.*, 2005) and studies with physical phantoms designed to mimic biological white matter (Campbell *et al.*, 2005; Perrin *et al.*, 2005). In addition, Bürgel *et al.* (2006) have generated maps of the routes of a number of fasciculi, based on postmortem histology, for comparison with tractography results. We saw, in Fig. 5.8, evidence of qualitative agreement between the *q*-ball ODF and fasciculus crossing information derived invasively; and in a similar way Schmahmann *et al.* (2007) demonstrated a very respectable agreement between DSI-based tractography and histology in the monkey brain. The matches are still far from perfect, however; and are often demonstrated in rather idealised conditions. The DSI scan used by Schmahmann *et al.*, for example, was performed on a 4.7 T system and



**Figure 5.11:** Thalamic parcellation using probabilistic tractography. Dividing the human brain into major cortical regions (a) and colour coding thalamic voxels according to their most probable projection into cortex using tractography (c,d) yields results broadly in good agreement with information obtained using invasive methods (b). Reproduced from Behrens *et al.* (2003a).



**Figure 5.12:** Kissing and crossing fibre architectures. In **(a)** the two fibre populations bend away from one another, whilst in **(b)** they cross or interdigitate. Since the angular information intrinsic to each of these scenarios is very similar, it is hard to tell them apart from their ODFs.



took 25 hours to complete. Such protocols are clearly useless to the clinician.

## 5.6 Summary

In this chapter we have attempted to provide a sense of the spectrum of extant approaches to the fibre tracking problem. We have focussed on giving a sense of the breadth of the alternatives, to avoid provoking informational indigestion in the reader (or the author), and have therefore omitted one or two notable techniques due to their similarity to other methods. It should be evident that the range of proposed solutions is wide, although they differ with respect to a fairly small number of core principles. Streamline-based tracking methods are the most widespread, but the model of orientation density is an important factor. There is, as yet, no clear reason to use one particular technique over all others, and studies based on tractography would be well advised to justify their choice of algorithm according to the nature of their aims.

# Neighbourhood tractography

IF A BREAKDOWN of communication between regions of the brain is indeed a significant factor in cognitive impairments, as the disconnection hypothesis posits, then dMRI and tractography are surely apt for studying this kind of pathology. But for the potential of the technique to be fully realised, its ability to provide proxy measures for white matter integrity such as FA will not be sufficient—it is also important that robust methods exist to compare such measures between normal and abnormal conditions, and to spatially localise any reproducible differences whenever possible.

In this chapter, we review the methods that have been applied to the localised study of white matter with dMRI, and describe a novel and automated approach to the issue. Our method treats segmentation as an *a posteriori* tract matching problem. We define a reference tract in a single brain volume, and then use a tract similarity measure to select from a number of candidate tract segmentations based on their topological resemblance to the reference. We demonstrate that this approach improves the consistency of segmentation results in a group of healthy young volunteers, thus reducing the impact of one source of within-group variance on anisotropy measurements.

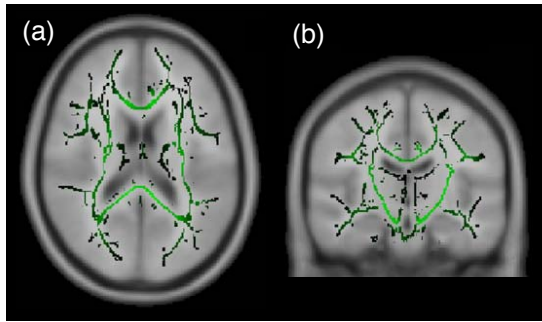
## 6.1 Group comparison in white matter

Approaches to the identification and localisation of systematic differences between the white matter of two or more populations fall into two broad categories. One can either use a technique that is itself capable of highlighting local regions where differences are focussed; or hypothesise where such regions may be, *a priori*, and then study those target areas specifically. Tractography, when applied as a segmentation technique, provides a white matter-specific tool for implementing the latter approach. However, we begin here by examining those methods which work with the whole brain.

Voxel-Based Morphometry (VBM) is a whole brain technique which was originally conceived to find areas of structural difference between groups (Wright *et al.*, 1995), but has since been generalised to voxelwise comparison of many types of medical image data (Ashburner & Friston, 2000; Good *et al.*, 2001). Uptake of the technique in the clinical dMRI literature has been significant over the past few years, particularly in the study of schizophrenia, whose effect on the brain is thought to be diffuse (e.g. Ardekani *et al.*, 2003; Burns *et al.*, 2003; Park *et al.*, 2004). In these applications, the analysis is typically performed on FA maps, and the comparison is between patient and control groups.

A typical VBM pipeline involves spatially normalising the set of images by transforming them into a common space, filtering the normalised images with a Gaussian smoothing kernel, and then performing a statistical comparison of intensity at each voxel. The smoothing process confers several benefits for both sensitivity and specificity: it improves the signal to noise ratio in the data; it makes correction for multiple comparisons less onerous<sup>a</sup>; and it helps to avoid

<sup>a</sup>Smoothing with a Gaussian kernel allows one to use the theory of Gaussian random fields to perform a multiple comparisons correction which is less conservative than a Bonferroni-type correction. However, the extent to which



**Figure 6.1:** A typical FA skeleton created by the TBSS technique with data from 10 healthy subjects. The skeleton and underlying standard brain image are shown in axial (a) and coronal (b) planes. Reproduced from Smith *et al.* (2006).

false positives due to misregistration between images (Ashburner & Friston, 2001). However, the choice of filter size is a problematic issue. Ideally the full width at half maximum (FWHM) for the filter should be approximately equal to the size of features of interest, but the spatial extent of population differences in anisotropy is usually not known in advance; and no other principled method for choosing the filter size has been established for dMRI data. As a result, kernels with FWHMs of between 3 mm and 16 mm have been applied to FA maps, usually without any explicit justification; but it has been shown that the eventual conclusion drawn from a data set can depend directly on this choice (Jones *et al.*, 2005a). Moreover, even when smoothing is applied, spurious “results” due to registration errors persist to some extent. Unfortunately, these two drawbacks represent significant limitations to the method.

The difficulty of choosing a filter FWHM illustrates an important general point. If one is interested in contrasting one group of subjects against another, at whatever scale, one must compare like with like. Consistency is a crucial prerequisite. But for findings to be meaningful and reproducible, the techniques they employ must additionally be robust. A strong dependency of a result on the value of any methodological parameter is strongly undesirable, particularly if there is no principled way to choose that parameter.

Concern over the limitations of vbm, as well as its lack of specificity to diffusion data, motivated the recent development of another whole brain method called tract-based spatial statistics (TBSS; see Smith *et al.*, 2006), which is just beginning to be applied to clinical data (Anjari *et al.*, 2007; Kochunov *et al.*, 2007; Rouw & Scholte, 2007).

The key innovation of the TBSS method is its use of an anisotropy “skeleton”—a ridge of locally maximal FA running through the brain’s white matter structures—to establish voxel homology for comparison (see Fig. 6.1). This approach allows one to perform voxelwise statistics without relying on the accuracy of image registration methods alone. By focussing on areas of the brain which can be confidently classified as white matter, the method additionally reduces the number of comparisons that need to be performed, thus improving the statistical power of any given data set. It is assumed that local variation in anisotropy *across* a tract is entirely attributable to partial volume effects.

Given a suitably preprocessed dMRI data set, the full TBSS pipeline consists of the following four steps.

1. Perform nonlinear registration of each individual FA map to the most typical one—that is, the one requiring the smallest average displacement. Resample each registered map to a standard resolution of  $1 \times 1 \times 1$  mm.
2. Average the registered maps and generate an FA skeleton for this average map, by finding the voxel with maximal FA along lines perpendicular to local tract directions. Threshold this skeleton at an FA value of around 0.2 to 0.3.
3. Search each individual FA map for locally maximal values in the same way, and project each separate skeleton onto the average one, thus establishing voxel homology.
4. Perform voxelwise comparisons within the skeleton.

the assumptions of this approach are met in FA images may be limited (for details see Jones *et al.*, 2005a).



Note that there is no direct spatial smoothing involved in this process—although the average  $FA$  map used for step 2 will be implicitly smoothed to some degree by registration inaccuracies—so the problem of choosing a filter width does not occur.

The capacity of whole brain analysis techniques like  $vbm$  and  $tbss$  to obviate the need for predefined brain areas of interest can be an invaluable one, particularly for exploratory studies where detailed *a priori* information is simply not available. However, just as the increased specificity of  $tbss$  provides gains in statistical power over  $vbm$ , the advantage of making more detailed hypotheses is that subtler effects can be found more easily. Hence, when prior information regarding the likely location of interesting effects is at hand, methods for studying a particular fasciculus come into their own.

## 6.2 Tract-specific comparison

Perhaps the simplest method for searching for tract-specific differences between populations involves manually superimposing regions of interest ( $ROIs$ ) with fixed dimensions onto an  $MRI$  image with high grey matter–white matter contrast. Indices of white matter integrity such as  $FA$  can then be averaged within these regions and compared between the subject groups of interest. This practice was employed in many of the first clinical comparative studies that used  $DTI$  (e.g. Ellis *et al.*, 1999; Jones *et al.*, 1999), and it is still far from obsolete.

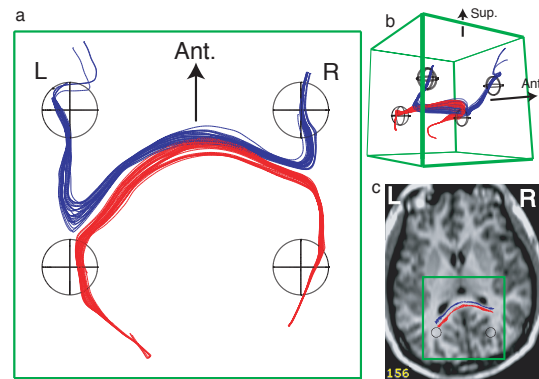
Manual placement of  $ROIs$  has several advantages, many of which stem from its simplicity. Computational run time is trivial. There is no complex relationship between the “original” (i.e. native space)  $FA$  data and those values which are used for comparison, as there is in the common-space comparisons of  $vbm$  and  $tbss$ . Depending on the level of specificity of the hypothesis, it may not be necessary to find congruent regions in each brain—placing an  $ROI$  in the correct structure with high confidence is often sufficient. Smoothing and multiple comparisons correction are typically minimal or unnecessary; and averaging within each region reduces the effects of noise. On the other hand, the method is not without its limitations. The choice of  $ROI$  size is arbitrary—too large and partial volume effects will be considerable; too small and noise will be a problem—although the relationship between this choice and the results is likely to be less complex than that of  $vbm$ ’s filter width. Study of any given fasciculus is limited to a very small and arbitrary portion of its length, and the region of the tract in which the  $ROI$  is placed will tend to depend on where it is most easy to fit it. Consequently, some tracts may be wholly excluded from this type of study because they are never wide enough to receive an  $ROI$  of the chosen size.

Probably the biggest drawback of manual  $ROI$ -based comparison, though, is its subjectivity. Although it requires almost no computational effort, the number of man-hours required to place the required regions in a typical data set is considerable; and it is hard to justify the particular placement choices made by any given observer. The related approach of segmenting an entire white matter structure by hand is not only time consuming; it is also extremely difficult to do well, since tracts are three-dimensional and can have highly irregular shapes.

Tractography provides an alternative. It is both objective and specific to white matter; and it lends itself directly to the segmentation of whole tracts, thus minimising noise issues whilst still focussing on a single fasciculus. The complex shape of white matter structures is not a problem. Of course, the caveats that apply to fibre tracking in general also apply to this application, but as problems such as crossing fibre degeneracy are handled better by new algorithms, so segmentation accuracy can be expected to improve.

The problem is then one of choosing seed points. Since the output of tractography algorithms is usually highly sensitive to the particular choice of starting location, care must be taken to ensure that a study is truly comparing like with like. Indeed, however sophisticated tractography algorithms become, the question of how best to initialise the tract reconstruction process is likely to remain an important one, especially in segmentation applications where consistency is important. Nevertheless, relatively little work has been done to date towards a principled and practical approach to seed point placement. Seeds are sometimes placed by hand, but average  $FA$  measured in tracts segmented in this way has been shown to vary quite widely between observers, and particularly between scans, even for a single subject (Ciccarelli

**Figure 6.2:** The two regions of interest constraint can separate disparate tracts which pass close to one another, in this case in the corpus callosum splenium. **(a,b)** Two and three-dimensional illustrations of the two tract trajectories, the former in axial projection. **(c)** The location of subfigure (a) in the brain. Reproduced from Conturo *et al.* (1999).



*et al.*, 2003a). Some of this variation will be due to noise, but subjectivity remains a major confound if one is looking for group differences.

Seed points can be placed in some standard space and then transferred to each subject's native space using an image registration algorithm. The assumption is that if the transformation between spaces is accurate then the seed points are congruous, and so the same fasciculus will be segmented in each brain volume. There is a very real risk, however, that registration errors and anatomical variation between subjects will make this assumption unsafe. Performance can be improved by weighting the registration cost function so as to maximise the importance of aligning white matter regions well, but we will see that this approach, which we will refer to as the *registration method*, still has problems. On the other hand, it has the advantage of being semiautomatic, and so reducing the effect of observer bias—although the choice of registration algorithm and its parameters will of course affect the outcome.

Choosing a single seed point *a priori* is not the only way to use tractography for white matter segmentation. One could instead seed at a number of voxels, which raises the questions of which seeds in particular to use and what to do with the multiple tracts that result.

There are several answers to these questions available in the literature, and the spectrum of supported responses is still tending to enlarge rather than shrink. Probably the most well established method is to constrain the tractography algorithm so that all reconstructed pathways must pass through two or more “waypoint” rois (Conturo *et al.*, 1999). The use of this constraint is illustrated in Fig. 6.2. Seeding near the middle of the corpus callosum splenium, in this case, may produce a streamline that projects posteriorly into occipital cortex (coloured red) or one that projects anteriorly (coloured blue). Since the pathways run very close to one another, the reconstruction will be highly sensitive to the exact location of the seed point. However, if one seeds in a number of locations in the splenium and then retains only those streamlines which pass through the two posterior rois, it is assured that the anterior projection will be ignored. Although these rois are often placed by hand or using registration-based transformation, they can often be far wider than the tract of interest—unlike regions used directly for segmentation—and their exact placement may therefore not be crucial. This kind of method is applicable to any streamline-based tractography algorithm, deterministic or probabilistic; and has been applied to group tractography (Abe *et al.*, 2004). There are some limitations, however, which we will discuss later in the chapter.

Another alternative is to seed throughout the brain. The advantages and disadvantages of this are obvious: on the one hand, choosing seed points is no longer an issue; on the other, a large proportion of the results are irrelevant to the study of any given fasciculus. A two roi constraint can be applied, or one can use clustering techniques to divide up a brainful of streamlines into related bundles. Various distance metrics have been proposed for streamline clustering (Brun *et al.*, 2004; Corouge *et al.*, 2006; Maddah *et al.*, 2005; O'Donnell & Westin, 2005), but one has still to identify the bundle or bundles of interest, and the general approach is not directly applicable to probabilistic tract representations.

The remainder of this chapter is dedicated to describing a novel perspective on the seeding problem, and applying it to some real-world data. We then compare the new method, which we refer to as *neighbourhood tractography* (NT), with region of interest-based alternatives.

## 6.3 Similarity and matching

Rather than modify tractography output to suit a particular criterion, the aim of the following work is to improve the consistency of tractography-based segmentation in group data by refining the initialisation of the algorithm; i.e. the seed point. In order to eliminate observer subjectivity, the method is to be automated. Our approach is to choose, from a group of “candidate” seed points, that point which produces the best output. In order to quantitatively define what constitutes “good” or “correct” output, we develop a novel tract similarity measure, based on the shape and length of two tracts being compared (first described in Clayden *et al.*, 2006a). To validate the measure, and demonstrate that it provides useful information, we use it to quantify similarity between independently generated comparable and disparate tracts in a group of volunteers. Finally, we define a series of reference tracts, and apply the measure to the problem of consistent seed point placement across this subject group, and show that the set of tracts thus derived are more visually similar to one another than the set produced by the registration method (cf. Clayden *et al.*, 2006b, from which Figs 6.3–6.8 are taken).

Since there is a diverse array of tractography algorithms available, and studies may wish to use different algorithms depending on the nature of the problem or hypothesis that they are working on, it is desirable that the process of tract matching be as independent as possible of the choice of algorithm. However, different algorithms produce different tract representations, as we saw in chapter 5, which creates a problem when we want to compare them using a single method. The solution is to use a common representation for all tracts, for the purposes of matching only. The “field of connection likelihoods” representation of a tract that is natural for probabilistic algorithms such as BEDPOST/ProbTrack can easily be generated by spatial discretisation of a deterministic streamline, so for the purposes of our tract similarity measure, we will assume that the tractography algorithm takes as input a single seed point, and produces voxelised, quantitative output. Hence we can define a tract  $r$  as the ordered pair

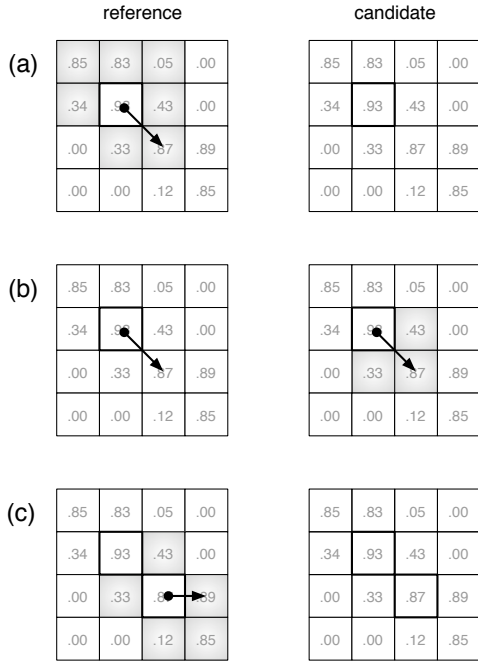
$$r = (\mathbf{a}_r, \phi_r(\mathbf{x})), \quad (6.1)$$

where  $\phi_r(\mathbf{x})$  is a discrete scalar field denoting the likelihood of a path from the seed point,  $\mathbf{a}_r$ , running through the voxel at location  $\mathbf{x}$  in the native acquisition space of the subject. These two data elements are tied together because they represent both the input and output of the tractography algorithm. If  $\mathbf{a}_r$  changes, then  $\phi_r$  will change too.

We will work on the principle that the characteristics of interest when comparing white matter tracts are length and shape. That is, if two tracts have the same shape *and* have the same length, then they are considered identical. For the purposes of comparison, we will make a distinction between reference and candidate tracts. There is no structural difference between the two, with both having the form given in Eq. (6.1), but similarity is always calculated for a candidate tract relative to a reference tract, rather than vice versa.

The following algorithm, which is based on a simplification and specialization of a general curve alignment algorithm (Sebastian *et al.*, 2003), provides sensitivity to the shapes of both the reference tract,  $r$ , and the candidate tract,  $c$ . Its output also depends on the length of the shorter of the two tracts. It moves along the two tracts simultaneously, voxel by voxel, finding a maximum likelihood pathway through the data,  $\phi_r$  and  $\phi_c$ , subject to certain path direction constraints. The output of the algorithm is a scalar value,  $\sigma(r, c)$ . The calculation is asymmetric, so that in general,  $\sigma(r, c) \neq \sigma(c, r)$ . The algorithm tacitly assumes that the seed points are equivalently located in the two tracts.

1. Initialise two sets of visited voxel locations,  $V_r$  and  $V_c$ , to the empty set.
2. Set tract pointers to the seed point location in each tract.
3. Add the current pointer position in the reference tract to the set  $V_r$ , and the position in the candidate tract to  $V_c$ .
4. Check the voxel values, from the field of connection likelihoods,  $\phi_r$ , of the 26 voxels forming a cube around the current pointer location in the reference tract, and choose the largest valued neighbouring voxel not in  $V_r$ . Note the step vector,  $\mathbf{v}_r$ , required to move to this new location.



**Figure 6.3:** Two-dimensional illustration of the shape similarity algorithm, as applied to two identical tracts. In (a), the boxes with bold borders represent the starting point, which has been marked visited. The shaded voxels in the reference tract indicate those nonzero, unvisited locations that the algorithm may legally move into, and the line represents the chosen step vector, from the current pointer location (circle head) to the next location (arrow head). In (b), movement in the candidate tract is restricted to those voxels whose angle from the chosen step direction in the reference tract is less than 90°. Since the voxel values are identical, the same direction is chosen. In (c), the next step in the reference tract cannot be back to the previous pointer location, since it is marked visited. In each diagram, numbers represent connection likelihood values at each voxel.

5. Prohibiting movement at any angle greater than or equal to 90° from the chosen step direction in the reference tract, find the largest valued neighbour to the pointer in the candidate tract that is not in  $V_c$ . Note the step vector used here,  $\mathbf{v}_c$ .
6. Add the normalised inner product of the two step vectors to the result,  $\sigma(r, c)$ .
7. Move in the directions of the chosen steps and update the pointers in each tract.
8. Return to step 3, and repeat until there are no unvisited, nonzero voxels adjacent to one of the pointers. At this point, the algorithm has followed the reference tract to its end in one direction.
9. Return to step 2, and repeat until there are no unvisited, nonzero voxels adjacent to one of the starting points. The algorithm has now followed the reference tract to its end in all directions.

The normalised inner product calculated in step 6 is given by

$$\frac{\mathbf{v}_r \cdot \mathbf{v}_c}{\|\mathbf{v}_r\| \|\mathbf{v}_c\|}, \quad (6.2)$$

which is equivalent to the cosine of the angle between the two step vectors. The formulation of step 5 may seem to be excessively restrictive, but it simply ensures that the result is not undervalued due to the pointers drifting in opposite directions along the tract. This is an important issue because seed points are rarely placed at tract extremities—since such areas tend to be associated with high directional uncertainty—and so traversal away from the seed point can usually be in two, almost equally likely, directions. Note that there is no angle restriction in step 4.

The value of the  $\sigma$  function is translation invariant; but because we compare the local absolute directions of the tracts relative to the dMRI acquisition coordinate system, rather than curvature, it is not rotation invariant. This is desirable, since we do not want to produce spurious matches between rotationally symmetric tracts such as the corpus callosum genu and splenium, or bilateral pairs.

### 6.3.1 The reduced tract

Tract data of the form given by Eq. (6.1) are not constrained to be a single voxel wide, and in general they will not be. Moreover, since the algorithm ceases stepping through the data when either tract terminates, the exact path taken through a reference tract can vary, and may be different during comparisons with different candidate tracts. This makes establishing an upper bound on the value of  $\sigma(r, c)$  extremely difficult.

In order to alleviate this problem, we define a reduced version of the tract  $r$  to include that subset of the nonzero data in  $\phi_r$  which is visited during the comparison of  $r$  with itself, a process that is illustrated, for a two-dimensional case, in Fig. 6.3. Parts (a) and (c) of the figure represent two consecutive iterations of step 4 of the algorithm, and part (b) illustrates step 5. The shaded squares in the figure represent those voxels that the algorithm is allowed to move into, and the boxes with bold borders indicate visited voxels. After this calculation of  $\sigma(r, r)$ , the reduced tract,  $\tilde{r}$ , is defined as

$$\mathbf{a}_{\tilde{r}} = \mathbf{a}_r \quad \phi_{\tilde{r}}(\mathbf{x}) = \begin{cases} \phi_r(\mathbf{x}) & \text{if } \mathbf{x} \in V_r \\ 0 & \text{otherwise,} \end{cases} \quad (6.3)$$

where  $V_r$  is the set of visited voxel vectors calculated by the algorithm above. While  $r$  and  $\tilde{r}$  are generally not identical, they are equivalent to the  $\sigma$  function in the sense that

$$\sigma(r, r) = \sigma(\tilde{r}, \tilde{r}) = \sigma(r, \tilde{r}) = \sigma(\tilde{r}, r), \quad (6.4)$$

because all voxel locations whose data value is nonzero in  $r$  but not in  $\tilde{r}$  are never visited. It must be remembered here that the tract data  $r$  includes the seed point,  $\mathbf{a}_r$ , since this property will not hold if the same voxel data but different seed points were to be passed to the  $\sigma$  function.

When comparing a tract to itself the inner product calculated in step 6 of the algorithm will always be unity, and so the algorithm is merely counting the number of steps taken. Thus, the value of  $\sigma(r, r)$  is exactly equal to the number of nonzero voxels in  $\tilde{r}$  (excluding the seed point), and since each nonzero voxel can be visited at most once, producing a maximum score contribution of one, we can establish the bounds

$$0 \leq \sigma(\tilde{r}, c) \leq \sigma(r, r) \quad \forall c. \quad (6.5)$$

The restriction that the pointer in the candidate tract can never move in a direction opposite to the reference tract ensures that all inner products are positive, and this fixes the lower bound in Eq. (6.5) at 0. Equivalently,  $0 \leq \sigma(r, \tilde{c}) \leq \sigma(c, c)$  for any  $r$ .

### 6.3.2 A similarity measure

Using the tract comparison algorithm described above, we now develop measures of shape and length similarity, and then combine them together to form an overall similarity score.

We first approximate the length,  $L_r$ , of tract  $r$  as the number of voxels visited when it is compared to itself, excluding the seed point, which is given by

$$L_r \equiv \sigma(r, r). \quad (6.6)$$

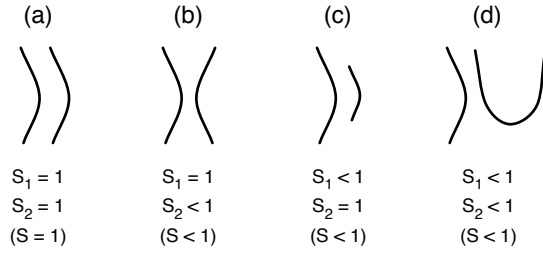
This length value is unchanged in the reduced tract,  $\tilde{r}$ , as shown by Eq. (6.4). Note that when comparing a tract to itself, shape is irrelevant because the local directionality of the reference and candidate tracts is always the same. If there are no nonzero voxels adjacent to the seed point, the data represents a “point tract”, with length zero.

Given the definition of length in Eq. (6.6), and having calculated its value for the reference and candidate tracts, we establish the similarity of these two numbers using the symmetric normalised difference given by

$$S_1(r, c) = 1 - \left| \frac{L_r - L_c}{L_r + L_c} \right| = \frac{2 \cdot \min\{L_r, L_c\}}{L_r + L_c} = S_1(c, r). \quad (6.7)$$

This measure has the value zero if either  $L_r$  or  $L_c$  is zero, and unity if the lengths are equal.

**Figure 6.4:** Qualitative demonstration of the effect on the two score components,  $S_1$  (length) and  $S_2$  (shape), of different types of relationship between the reference tract (fixed, and on the left in each case) and the candidate tract (variable, and on the right). The seed points are assumed to be in the centre of each tract throughout.



The other component of the similarity measure, the similarity in shape between the reference and candidate tracts, can be established using the asymmetric formulation

$$S_2(r, c) = \frac{\sigma(\tilde{r}, \tilde{c})}{\min\{L_r, L_c\}} \neq S_2(c, r). \quad (6.8)$$

The denominator in Eq. (6.8) removes the length dependence of the  $\sigma$  function. The bounds on the  $\sigma$  function that were established above ensure that the value of Eq. (6.8) is always in the interval  $[0, 1]$ .

Finally, the two score components given by Eqs (6.7) and (6.8) are combined to form the overall similarity score,

$$S(r, c) \equiv \sqrt{S_1(r, c) \cdot S_2(r, c)} = \sqrt{\frac{2 \cdot \sigma(\tilde{r}, \tilde{c})}{L_r + L_c}}, \quad (6.9)$$

the geometric mean of the two components. A higher value of Eq. (6.9) indicates a better match, and a lower value indicates a worse match. The score will be 1 if  $r$  and  $c$  are the same tract. It will be 0 if either  $r$  or  $c$  is a point tract. The geometric mean lends a far stronger influence to very small values in one score component than does the arithmetic mean when finding the “average” similarity of  $c$  to  $r$ , and in particular, if *either* score component is 0 then the overall score is also 0. This formulation emphasises that *both* length and curvature must be similar for the candidate tract to be considered a likely equivalent to the reference.

Fig. 6.4 shows four examples of tract pairs and their associated score components. In each case the reference tract is on the left, and the seed points are assumed to be placed exactly in the middle of each tract. These are idealised, and continuous rather than voxelised, tract curves; but they illustrate how the two score components will be affected in various scenarios. In (a), the candidate tract is identical to the reference tract. This is equivalent to the case in Fig. 6.3. In (b), the candidate is a reflected copy of the reference. Note that the shapes of these two curves are considered different. In (c), the candidate is a central segment from the reference, so the shape is considered identical, but the lengths differ. It should be noted that this case represents a truncation rather than a scaling of the reference tract, as the latter would not produce an  $S_2$  score of 1. Finally, in (d), the tracts are different in both shape and length.

## 6.4 Validation and application

Six normal volunteers (2 male, 4 female; mean age  $27 \pm 3.4$  years) were recruited for this study. Each subject underwent a dMRI protocol on a GE Signa LX 1.5 T clinical scanner, consisting of a single-shot spin-echo echo-planar imaging sequence with 51 noncollinear diffusion weighting gradient directions at a  $b$ -value of  $1000 \text{ s mm}^{-2}$ , and 3  $T_2$ -weighted scans. 48 contiguous axial slice locations were imaged, with a field of view of  $220 \times 220 \text{ mm}$ , and a slice thickness of 2.8 mm. The acquisition matrix was  $96 \times 96$  voxels in-plane, zero filled to  $128 \times 128$ .  $TR$  was 17 s per volume and  $TE$  was 94.3 ms.

In order to investigate the variation in similarity scores between acquisitions, 2 of the subjects were scanned twice, and 3 were scanned three times. Those subjects that went through the protocol three times were taken out of the scanner between the second and third acquisitions, and the slice locations were repositioned for the third acquisition without reference to those chosen for the first two.

The data were initially preprocessed to remove skull data and eddy current induced distortion effects from the images, using `fsl` tools. The underlying tractography algorithm used in this study was the `BEDPOST/ProbTrack` algorithm (Behrens *et al.*, 2003b). It should be remembered that the `BEDPOST` model of the `dmri` signal is a partial volume model assuming a single anisotropic diffusion direction at each voxel, and the measure of anisotropy it uses is the anisotropic volume fraction (`AVF`), rather than the more common, diffusion tensor-based fractional anisotropy (`FA`). However, the two measures are closely related.

The aim of our first experiment was to validate the similarity measure described above, by investigating whether the measure could differentiate between comparable and disparate tracts in the group of volunteers. A series of 8 seed points were placed in major white matter fasciculi on a Montréal Neurological Institute (`MNI`) standard brain (Evans *et al.*, 1993), and transferred to each subject's native space using the `FLIRT` registration algorithm (Jenkinson & Smith, 2001), with the `MNI` white matter map used as a weighting volume (Clayden *et al.*, 2005). The specific seed regions chosen were genu and splenium of corpus callosum (`CC`), right and left anterior limb of internal capsule (`ALIC`), right and left posterior limb of internal capsule (`PLIC`), and right and left sagittal stratum (`SS`). Whilst the accuracy of seed point placement using this registration method may be limited, it provides an independent mechanism for generating groups of tracts that can be expected to be more or less similar to one another. The `ProbTrack` tractography algorithm was run with each of these points as a seed, and similarity scores were calculated for various tract pair permutations. Comparisons between equivalent seed regions on the left and right of a single brain volume (e.g. left `ALIC` versus right `ALIC`) were labeled "bilateral", and all other comparisons within a single volume (e.g. left `ALIC` versus right `PLIC`) were labeled "nonbilateral". Comparisons across subjects for a single seed region (e.g. left `ALIC` in subject 1 versus left `ALIC` in subject 2) were labeled "intersubject"; and additional similarity scores were calculated between 1st and 2nd scans ("inter-NEX"<sup>b</sup>) and 2nd and 3rd scans ("interscan"), where available, within each subject and seed region. We expect that similarity scores will be lowest for the nonbilateral comparisons, and highest for the interscan and inter-NEX cases where the two tracts are from the same seed region and same subject. For every pair of tracts thus compared, similarity scores were calculated using each in turn as the reference tract.

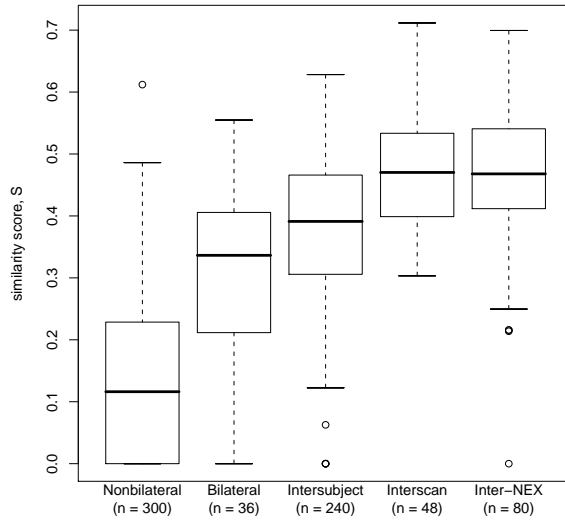
A second experiment was then performed, aimed at applying the similarity approach to the problem of improving the robustness of seed point placement across a group of scans. For each seed region, a representative reference tract was chosen from a single scan. For each other scan, a  $7 \times 7 \times 7$  cube of voxels around, and including, the voxel suggested by the registration method—hereafter the "original" seed point—for each fasciculus of interest were used as seed points for the tractography algorithm, except where the voxel `AVF` was less than 0.2, an empirically chosen threshold used to avoid seeding in cerebrospinal fluid or grey matter. The tract with the highest similarity score when compared to the relevant reference tract was then selected as the "best" tract from each brain volume. We refer to this technique as neighbourhood tractography.

In all of the experiments described above, reference and candidate tract data (i.e. the fields  $\phi_r$  and  $\phi_c$ ) were thresholded at the 1% level before similarity scores were calculated. This was done to avoid inclusion of very low confidence paths in the comparisons.

Fig. 6.5 shows the results of the first experiment as a box-and-whisker plot. The mean ( $\pm$  one standard deviation) similarity score for each group of tract comparisons was 0.14 ( $\pm 0.13$ ) for nonbilateral, 0.31 ( $\pm 0.13$ ) for bilateral, 0.38 ( $\pm 0.12$ ) for intersubject, 0.47 ( $\pm 0.09$ ) for interscan, and 0.46 ( $\pm 0.12$ ) for inter-NEX. Two sample, two tailed *t*-tests showed significant differences between nonbilateral and bilateral scores ( $P < 10^{-9}$ ), between bilateral and intersubject scores ( $P = 0.005$ ), and between intersubject and interscan scores ( $P < 10^{-6}$ ). There was no significant difference between interscan and inter-NEX similarity scores ( $P = 0.89$ ).

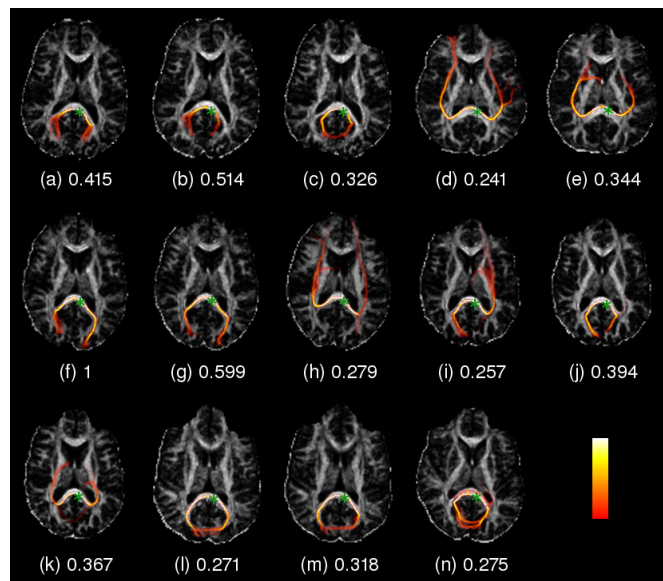
Results from the second experiment are shown visually in Figs 6.6 and 6.7. The correspondence between the letters labeling each subfigure and the different scans is shown in Table 6.1. Fig. 6.6 shows the tract fields produced by seeding `ProbTrack` at the original seed point in splenium of corpus callosum, and thresholding the results at the 1% level. This seed region

<sup>b</sup>The acronym `NEX`, for "number of excitations", is commonly used to denote the number of times an imaging sequence was applied to the subject.



**Figure 6.5:** Box-and-whisker plot showing the range of similarity scores for the five different categories of comparison in the first experiment. The thick horizontal line across each box represents the median, the box shows the interquartile range, the whiskers show the extent of the bulk of the data, and circles show outliers more than 1.5 interquartile ranges from the box. The  $n$  values indicate the number of scores making up the data for each plot. The data demonstrate appropriate score increases across the different test conditions, suggesting that the score provides meaningful and useful information.

**Figure 6.6:** Two-dimensional axial projections of the tracts generated by the ProbTrack algorithm using the original seed points chosen by the registration method, overlaid on AVF maps of the slice in plane with the seed in each case. White indicates high AVF and black low. In the tracts, yellow indicates high likelihood of connection to the seed point, and red low. The green stars indicate the seed point locations. The similarity score to the reference tract (f) is shown in each case.





Subject	Scan 1	Scan 2 (inter-NEX)	Scan 3 (interscan)
1	(a)	(b)	(c)
2	(d)	(e)	
3	(f)	(g)	(h)
4	(i)	(j)	(k)
5	(l)	(m)	
6	(n)		

**Table 6.1:** Correspondence between the different scans and the subfigure labels used in Figs 6.6 and 6.7.

Seed	MNI point	RM score mean	RM score s.d.	NT score mean	NT score s.d.
cc genu	(8,22,14)	0.488	0.056	0.597	0.018
cc splen.	(-6,40,14)	0.354	0.106	0.542	0.031
right ALIC	(18,10,6)	0.529	0.098	0.651	0.026
left ALIC	(-16,10,6)	0.463	0.099	0.644	0.027
right ss	(36,-54,10)	0.329	0.220	0.680	0.023
left ss	(-36,-54,10)	0.365	0.077	0.516	0.024
right PLIC	(22,-14,10)	0.405	0.096	0.570	0.030
left PLIC	(-22,-14,10)	0.444	0.054	0.594	0.025

**Table 6.2:** Mean and standard deviation of similarity scores for all tracts chosen by neighbourhood tractography (NT) in each of the 8 seed regions, determined from the 6 volunteers (14 scans). The means and standard deviations for tracts chosen by the registration method (RM) are given for comparison. The position of the seed point in MNI standard space is given in millimetres.

was chosen as the example because considerable variation in tract shape can be seen across the group: the resultant tracts demonstrate pathways running anterior (d, e, h, k), posterior (a–c, f, g, j, l–n) or both (i) from the edges of the corpus callosum itself (cf. Fig. 6.2). Fig. 6.7 shows the tracts chosen by the neighbourhood tractography approach, after the same 1% threshold has been applied. Both figures also show the similarity scores associated with each tract, using (f), which is the same in both cases, as the reference tract. In Fig. 6.7, similarity scores are necessarily greater than or equal to the corresponding score in Fig. 6.6, and only two tracts (i, l) remain that do not project in the posterior direction from the corpus callosum. These two tracts have the two lowest similarity scores in the figure. Tract (g), which has the highest score apart from the reference tract, is found in the same subject as the reference tract, so the fasciculus it represents is identical.

Fig. 6.8 shows examples of “reduced” reference and candidate tracts, in the sense described in §6.3.1. It shows how the reduction affects the tracts. In this case, the reference tract is simply slightly narrower than its unreduced equivalent, Fig. 6.6(f). The candidate tract is truncated at the edge of the splenium, where the unreduced version, Fig. 6.6(h), had an ambiguous branch.

The mean and standard deviation of the similarity scores for the tracts chosen before and after applying neighbourhood tractography for each seed region, across all subjects and acquisitions, are given in Table 6.2. The figures for the “best” tracts—as chosen by neighbourhood tractography—represent narrow and seed-specific score distributions, whose coefficients of variation (cvs, the standard deviations divided by the means) are in the range 3.0–5.7%. By comparison, the original scores, generated by the registration method, are invariably lower with wider standard deviations. Their cvs are in the range 11.5–66.9%.

In the second experiment, 63% of seed points chosen by neighbourhood tractography were not more than 2 voxels from the original seed point in any direction. This proportion is high enough to suggest that a  $7 \times 7 \times 7$  search neighbourhood is generally sufficiently large. Run time for the method might be expected to increase with the cube of the neighbourhood width, although in practice it may be that more of the extra seed points are rejected by the anisotropy threshold, and so the scale factor may be a little less.

## 6.5 How many seeds?

To date, it has not been explicitly shown that tractography-based segmentation using a single seed point cannot yield consistent results, and we see no reason that this should be the case, particularly as the sophistication of tractography algorithms continues to improve. However it is certainly true, as we saw in §6.2, that choosing a single point so as to obtain useful results is hard. We have described here a method which emulates a process for selecting an appropriate seed that might be used by a human observer: the expected topology of the tract is clearly defined—in this case in terms of a reference tract—and then we try seeding at several plausible locations until a good match is found. Unlike the human observer, though, the algorithm is completely consistent in its assessment of candidate tracts and has no difficulties working in three dimensions. The selection process is also far faster using the algorithm: comparing each candidate tract with the reference tract takes only a second or so on a typical workstation.

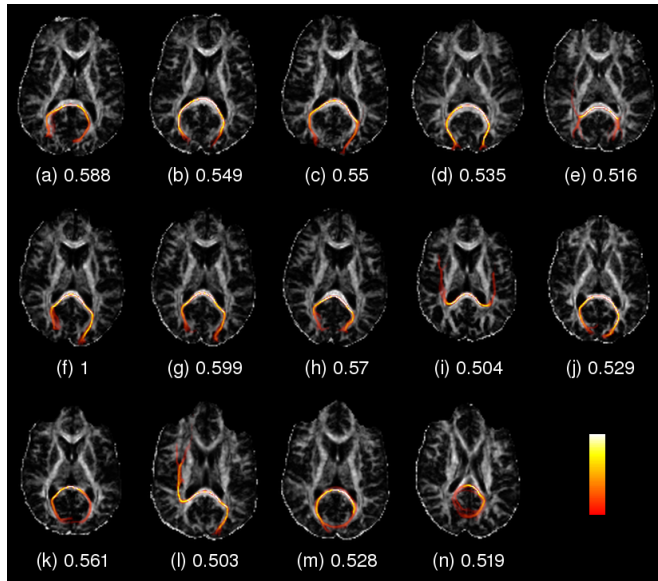
The relationship between the neighbourhood tractography method, as we have presented it here, and region of interest-based methods is simple to explain. Let us assume that a plausible but suboptimal seed point has been selected in the native space of the subject using the registration method. This seed point could then be used directly for tractography, or one could grow an ROI around it and seed at every point therein, possibly subject to an anisotropy threshold. Neighbourhood tractography then performs all-but-one rejection of these seed points based on *a posteriori* tract similarity, whereas a multiple ROI method would combine results from the whole seed region, subject to the waypoint constraint on individual streamlines. Another option is to simply retain all the results with no constraints: this is a single ROI method. Selection between these strategies is therefore partly a question of deciding how many seed points should be used to generate the final tract representation.

It should be noted that the two ROI constraint, as demonstrated by Fig. 6.2, is effectively a modification of the tractography algorithm that it is applied to. A deterministic streamlining algorithm, thus modified, will return either a streamline passing through the seed point and the waypoint regions, or nothing. In the probabilistic case, the effect is to add an extra conditional dependency on the ROI locations to the connection likelihoods, making it somewhat more difficult to interpret the results.

Fig. 6.9 shows the results of applying the different strategies in a single scan, corresponding to Fig. 6.6(c), using 5000 streamline samples per seed point in each case. If the reference tract in Fig. 6.6(f) is taken to represent the pathway we are attempting to find in this scan, then Fig. 6.9(a) surely shows poor correspondence. Its projection into cortex is further anterior than it ought to be, and the reconstructed tract appears to cross the interhemispheric plane posterior to the splenium, which is definitely nonphysical. By contrast, tract (b) shows neither of these problems. It can be seen in the unthresholded version of this tract, (f), that a very small number of probabilistic streamlines (less than 1%) do project anteriorly from the splenium, but the main trajectory of the tract is consistent with the reference.

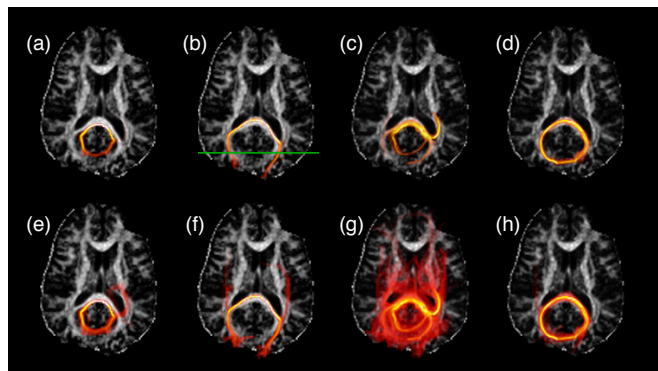
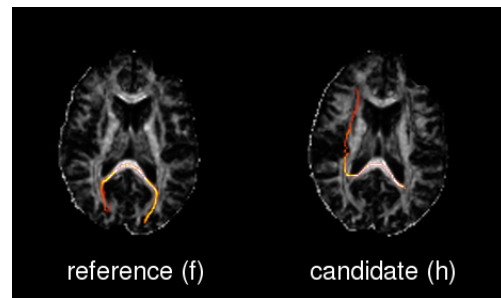
Full seeding in the  $7 \times 7 \times 7$  voxel neighbourhood yields tract (g), which is very widely spread out and heavily affected by thresholding—compare tract (c). These two effects are related. A very large number of probabilistic streamlines are generated in this case (570,000 in total), but since they are very widely spread out no more than 80,710 pass through any single voxel in the brain. As a result, only 5.2% of nonzero voxels from (g) survive the threshold, compared to 21.5% from (f). Worse, the threshold scales with the volume of the neighbourhood, but the visit counts at each voxel do not keep pace (see Fig. 6.10), making the application of this kind of threshold undesirably sensitive to neighbourhood size. This effect is even more pronounced if an anisotropy threshold is not used to cull unpromising seeds from the seeding region. It seems, then, that tract (c) should be treated with caution, while (g) is too nonspecific to be of much use. Nevertheless, by carefully placing a smaller ROI well within the tract, a single seeding region can be practical (Kanaan *et al.*, 2006).

Finally, tracts (d) and (h) show the result of using two constraint ROIs in addition to the seeding ROI. In order to yield a tract as similar as possible to (b), we used information from the latter to place the constraint regions. The centres of the waypoint ROIs in the left and right hemispheres were placed at the locations with the greatest voxel value, in (b), within a plane normal to the anterior–posterior axis that is shown in green. The size of the constraint

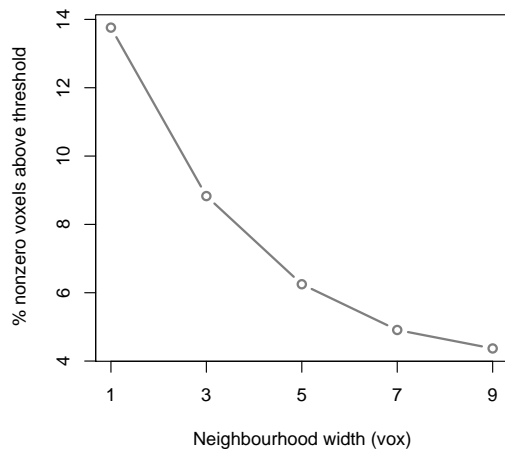


**Figure 6.7:** Projections of the tracts chosen as the “best” (highest similarity to the reference tract), using a  $7 \times 7 \times 7$  seeding neighbourhood around the original seed point. Individual similarity scores are also shown. Tract (f) is the reference tract.

**Figure 6.8:** Examples of reduced reference and candidate tracts, produced from two of the unreduced tracts shown in Fig. 6.6.



**Figure 6.9:** Illustrative results using various seeding strategies for the corpus callosum splenium in a single subject, ignoring potential seed points with AVF of less than 0.2 throughout. (a,e) Single seed point placed using the registration method. (b,f) Single seed point chosen using neighbourhood tractography. (c,g) Full seeding throughout the neighbourhood, with additive combination of results. (d,h) Full seeding, but with additional bilateral waypoint ROIs placed posterior to the splenium. (a–d) are thresholded at the 1% level; (e–h) are unthresholded. The colour scale is not consistent in meaning between subfigures.



**Figure 6.10:** The effect of thresholding on connection likelihood when using a single ROI strategy. The impact of the threshold, expressed as a fixed proportion of the total number of streamlines initiated in each case, is more significant as the seeding ROI increases in size.

rois was  $7 \times 7 \times 7$  voxels, as with the seeding region. The effect of thresholding in this case is far more modest than without the constraints, but again the segmentation appears to be passing between hemispheres twice, suggesting that the use of two waypoint regions may not be sufficient to ensure a plausible segmentation. One can attempt to rectify this by adding more constraints—in a recent reproducibility study, up to five constraint rois of four different types were used to segment each of the fasciculi of interest (Heiervang *et al.*, 2006)—but all such restrictions are simultaneously absolute and independent of the actual data, and so risk undermining the very advantages of using tractography for segmentation. The logical limit of the process of adding constraints, after all, is the case in which tracts are simply outlined by hand.

It is interesting to note that even though the neighbourhood tractography result was used to inform the placement of waypoint rois, the result is very different in shape. The pathway segmented in (b) is present in (h), of course, but the projection is so “unlikely”, given the seed mask and waypoint constraints, that its end points in occipital cortex are absent from the thresholded version. Instead, the tract projection in (d) represents the most visited pathway from (g) that passes through the waypoints. It is very difficult to decide which of these results is more anatomically correct. Validation of tractography output is a complex issue in its own right, as we discussed in §5.5. However, we would certainly claim that the tracts shown in Fig. 6.7 are more similar to the reference tract—which is chosen for illustration—than those in Fig. 6.6. If the nature of the reference tract were later to be found to be inappropriate, it could be updated and neighbourhood tractography repeated without change. The method would then find the best match to the new reference tract.

Although we have focussed on probabilistic tractography in this comparison, our observations are just as valid in the deterministic—or maximum likelihood—case, where only one streamline is generated per seed point. However, a connection likelihood threshold would not then be relevant.

Multiple seed points for tractography need not necessarily be adjacent to one another in a neighbourhood. We have described, in separate work, an iterative approach in which each seed point is chosen from the tractography output for the previous seed such that it is as far as possible from its predecessor whilst still having high likelihood of connection to it (Clayden *et al.*, 2005). This can help the segmentation to recover, to some extent, from a poor starting seed point; but the choice of each seed point is not very strongly principled in that case, and the method is likely to be rather too scattershot for general purpose use.

Ultimately, the question of which seeding strategy generates the most useful segmentation of the fasciculus of interest will depend on the application. The all-but-one rejection approach of neighbourhood tractography offers the greatest specificity whilst providing a strong *preference*—as opposed to an enforced constraint—for tracts with topologies similar to the reference

tract. ROI methods, meanwhile, offer to segment the full width of the fasciculus if they are given a large enough seeding region. The latter may be desirable, or it may be problematic due to the larger expected role of partial volume effects. Of course, these two techniques are not mutually exclusive, and it may be that in some cases a combination is the most successful.

## 6.6 Evaluation of the similarity measure

While tract shape has been studied before (Batchelor *et al.*, 2006; Corouge *et al.*, 2004; Ding *et al.*, 2003), previous work has been aimed at modelling individual tracts, rather than doing pairwise similarity scoring. The kinds of tract characteristics that these previous studies have worked with, such as curvature and torsion, could in principle be applied to the tract matching problem; but as far as we are aware, the work described above represents the first actual attempt at using a quantitative tract similarity measure to improve segmentation consistency.

The results from our first experiment provide evidence that the similarity measure described above produces higher scores for a single seed region across a range of healthy subjects, than it does for a range of seed regions within a single subject, as demonstrated by higher intersubject than bilateral and nonbilateral similarity scores. Behaviour of this nature is clearly crucial for any tract similarity measure that is intended to be used as a basis for the identification of comparable tracts across a group of subjects. It is not surprising to find that comparisons between bilateral seed regions (such as left versus right ALIC) produce generally higher scores than other comparisons (such as left ALIC versus right PLIC), since comparable white matter fasciculi in the two hemispheres can be expected to have similar lengths and related shapes. Nevertheless, even the bilateral scores are significantly smaller than the intersubject scores.

The finding that interscan and inter-NEX scores are indistinguishable is an interesting one. It suggests that repositioning of the slice positions introduces no consistent bias to the results of the similarity measure, demonstrating a useful robustness to subtle changes in the slice locations. It is also reassuring to see that both these sets of scores are significantly higher than the intersubject scores, since the underlying fasciculi are *the same* across acquisitions, rather than merely comparable as they are in the intersubject case.

The narrowness of the score distributions for each seed point—as shown in Table 6.2—seems to indicate that the scoring algorithm is quite strongly influenced by the nature of the reference tract. This may be because the part of each tract near the seed point in each direction is relatively reproducible, whereas the spatially uncertain regions near the ends of tracts are very unlikely to produce a perfect match with the reference tract. The combination of these two factors may effectively impose reference tract-specific upper and lower score bounds.

A major advantage of the NT approach is that no spatial manipulation of each individual brain volume is required before tractography can be performed, and so potentially interesting anatomical variation across the group need not be averaged away or otherwise distorted. However, we have made no alterations or corrections for factors such as natural variation in brain size and shape, or head rotation; even for the purposes of tract comparison. In fact, a correction based on a transformation of the candidate tract into the space of the reference tract would have problems of its own, since interpolating the tract data could alter its structure in undesirable ways. For example, local duplication of voxel values—which would arise from a nearest neighbour interpolation scheme—would be strongly suboptimal for our similarity algorithm. The difficulty with registration also makes simple field-based distance measures such as the sum of squared differences between voxels highly problematic for comparing tracts.

Since differences in head rotation and head size between scans will have a complex, non-linear effect on the similarity measure, and may affect different tracts differently, it is not straightforward to establish the impact of these variates, nor to recommend upper bounds on acceptable rotations or scalings. Moreover, working with simulated data would add another image processing step, which may be a source of variance, and would introduce similar interpolation issues to a correction. However, interscan rotations for single subjects are present in our data set. Linear registrations between pairs of  $T_2$ -weighted images suggest that the median rotation between a subject's first scan and their third was  $1.5^\circ$  ( $4.3^\circ$  about the left–right axis,  $0.6^\circ$  about the anterior–posterior, and  $1.1^\circ$  about the superior–inferior). Hence, some

variance due to rotation is incorporated into the results from our first and second experiments; but it should be remembered that in the first experiment, inter-NEX and interscan scores were statistically indistinguishable, despite much smaller rotations in the former case (median of  $0.3^\circ$ ), suggesting a certain robustness to such effects.

The similarity measure described above aims to be relatively simple whilst capturing important characteristics of the two tracts that we wish to compare. This simplicity aids portability. Whilst probabilistic tractography algorithms tend to produce tract data of the form given by Eq. (6.1), some other approaches, particularly streamline-based algorithms, instead produce a single line of infinitesimal thickness through the seed point. In these cases, the principle of our similarity calculation would still be applicable, and in fact the method would become even simpler because there would no longer be any need to produce a reduced tract.

There is an obvious limitation of comparing shape at the voxel scale, which is that voxel sizes vary between data sets and have no intrinsic physiological significance at all. Moreover, since voxels are often not equal in width in all dimensions, a step of “one voxel” may represent a different real-world distance depending on the orientation of the step.

The main weakness of the similarity measure presented here is that the termination criterion in step 8 of the algorithm (see page 66) can be met prematurely if a local “loop” of relatively high valued voxels is encountered. This leads to underscoring or false negatives, and is likely to be at least a contributor to the problem of narrow score distributions for a particular reference tract, and the reason that tract (l) is *less* visually similar to (f) in Fig. 6.7 than in Fig. 6.6. That result, when taken in context with the rest of the data, suggests that while a high score seems to indicate a good match between tracts, a low score may not reliably indicate a bad match. Indeed it is plausible, even likely, that in some cases better matching tracts than those selected by this similarity measure were available but were underscored and therefore disregarded. This substantial issue could perhaps be alleviated by biasing the algorithm in favour of continuing in the same direction as its previous step, and introducing some fuzziness into the choice of local maximum voxel in step 4 of the algorithm. However, these changes would render the algorithm nondeterministic, and care would have to be taken to ensure that the maximum and minimum scores remain tractable. A different approach may be preferable.

## 6.7 The next step

This chapter began with a look at how segmentation and comparative analysis in white matter can be approached using either a tract-specific or a whole brain starting point, the choice between which should depend on the specificity of one’s hypothesis. We saw that segmentation using tractography is typically subjected to constraints based on a number of “waypoint” regions of interest, which are described in advance. This multiple ROI approach is one way to incorporate prior knowledge into the fibre tracking process, but it represents a hard constraint which complicates the interpretation of the resulting tract. The novelty of the neighbourhood tractography method, as an alternative approach, is that it allows prior information to be introduced in the form of a reference tract, but rather than constraining tractography directly, the reference is used to select from a number of candidate segmentations generated with different seed point initialisations.

We have demonstrated here that NT can work as intended, although we have also identified some shortcomings in the particular similarity measure that we used. As well as considering how these issues might be overcome, it is important to investigate how successful the method is when applied to data of clinical interest, where there may be more confounding factors than are present in scans from healthy young volunteers. The latter will be the focus of the next chapter.

# Applications

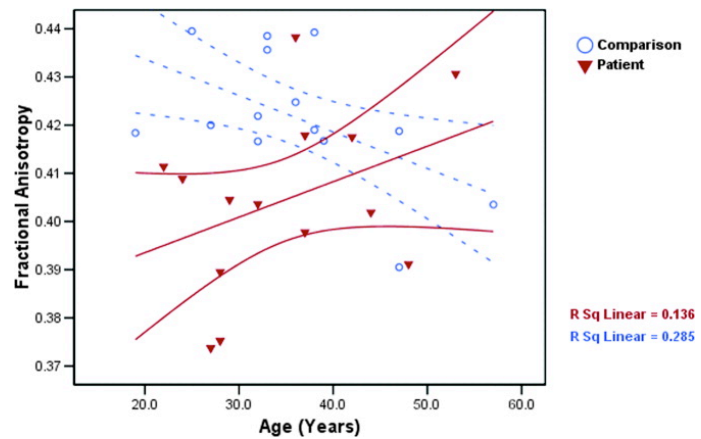
THE AIM of neighbourhood tractography is to facilitate comparative analysis between subject groups in clinical studies, in a tract-specific manner. The brains of unhealthy or aged individuals are, however, often substantially different from those of healthy young volunteer subjects. It is therefore important to confirm that the topological tract matching principle by which NT works remains valid in these cases. This chapter describes the application of NT to the clinical study of normal ageing and schizophrenia, and demonstrates that gains in segmentation consistency can be obtained even when the reference tract is drawn from a different population to the candidate tracts. The work described here was completed collaboratively with Jakub Piątkowski and Dr Susana Muñoz Maniega.

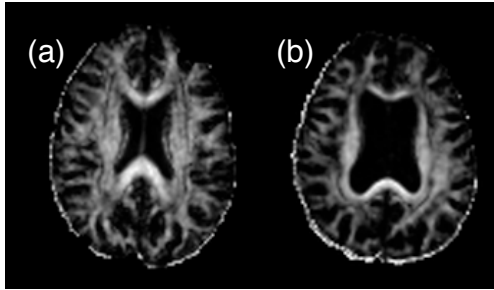
## 7.1 Tractography in the ageing brain

We discussed in §4.5 that normal ageing is a significant area of clinical interest in which dMRI has already begun to make a useful contribution. Early DTI studies of the effects of ageing on white matter, such as those by Pfefferbaum *et al.* (2000) and O'Sullivan *et al.* (2001), used manual segmentation of large white matter regions of interest, and demonstrated negative correlations between diffusion anisotropy and age. O'Sullivan *et al.* found a particularly strong effect in anterior white matter—a finding which has since been reproduced by Head *et al.* (2004); and for the corpus callosum genu in particular, by Abe *et al.* (2002). Kochunov *et al.* (2007) have additionally shown, using the TBSS technique, that FA in the genu shows a more robust association with other indices of structural health in the brain—such as average grey matter thickness—than does anisotropy in other white matter regions. Evidence for similar frontal effects in ageing monkeys has also been recently demonstrated (Makris *et al.*, 2007).

Considering this increasing body of evidence that suggests that dMRI-based indices such as FA may be useful for studying ageing, it is surprising that studies employing tractography-based

**Figure 7.1:** Relationships between age and FA in schizophrenics (red triangles) and healthy controls (blue circles). Points are averages over eight tracts.  $R^2$  values for linear fits are given in each case. Reproduced from Jones *et al.* (2006).





**Figure 7.2:** The brain of a young adult (a) differs most obviously from that of a healthy elderly subject (b) in the volume of the ventricles. This difference is clearly visible in these AVF maps.

segmentation for examining specific tracts appear to be almost nonexistent. Such tract-specific information, obtained in a more objective manner than is possible with manual segmentation, could be particularly helpful for confirming or contradicting the suggestion that frontal white matter decline is particularly marked during normal ageing. Jones *et al.* (2006) provided evidence that the relationship between anisotropy and age appears to be different in schizophrenic patients to controls in general (see Fig. 7.1), but their tract-specific measurements relate only to the effects of schizophrenia, and are therefore not especially helpful in understanding the impact of ageing in the healthy population.

Tractography in the aged brain encounters additional challenges, compared to similar tracking in young adults. Firstly, since anisotropy is generally lower, the level of uncertainty associated with dMRI estimates of fibre orientation can be expected to be higher. This may make consistent segmentation of particular tracts intrinsically more difficult. Secondly, the morphology of older brains usually differs from younger ones—in particular, grey matter volume tends to shrink and the CSF-filled ventricles become larger (see Fig. 7.2). If this effect turns out to be highly variable among a population of aged brains, then using a reference tract to guide tract matching may not be as reliable as in younger brains.

To test the performance of NT in an aged population, 27 healthy volunteers aged over 65 were subjected to a dMRI protocol using a single-shot spin-echo echo-planar imaging sequence with 64 noncollinear diffusion weighting gradient directions at a  $b$ -value of  $1000 \text{ s mm}^{-2}$ , and 7  $T_2$ -weighted scans. 53 contiguous axial slice locations were imaged, with a field of view of  $240 \times 240 \text{ mm}$ , and a slice thickness of 2.5 mm. The acquisition matrix was  $96 \times 96$  voxels in-plane, zero filled to  $128 \times 128$ . TR was 13.5 s per volume and TE was 75 ms. It should be noted that these parameters differ a little from those used for the study described in chapter 6, although all subsequent image preprocessing steps were the same.

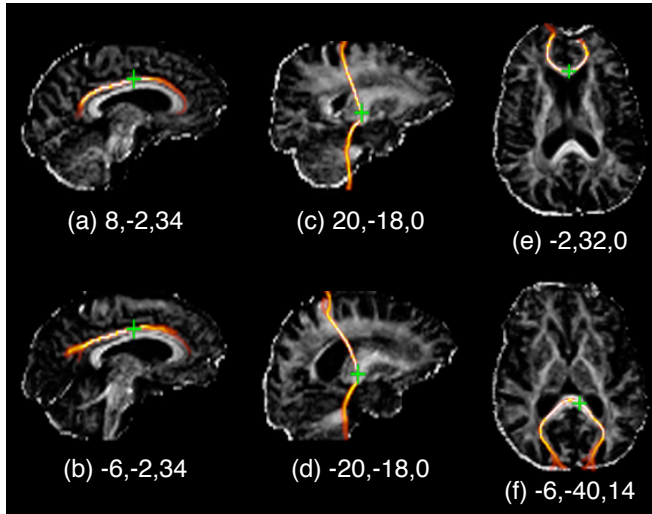
These data were acquired as part of a study called DELCORT, whose principal investigator is Dr Alasdair MacLullich, a Lecturer in Geriatric Medicine at the University of Edinburgh. All image preprocessing, tractography and reference tract selection for this section was carried out by Jakub Piątkowski, with assistance from Dr Mark Bastin and the author.

The fasciculi of interest that were used for testing NT in this aged cohort were the genu and splenium of the corpus callosum; the corticospinal tract (CST, left and right)<sup>a</sup>; and the cingulum bundle (CB, left and right). The registration method for seed point placement was used to transfer a single point for each tract from MNI standard space (Evans *et al.*, 1993) to each individual's brain. A reference tract was then selected by hand from the set of native space tracts, whenever an acceptable segmentation was available. In the corpus callosum genu, however, none of the tracts generated in this way was satisfactory, and so a seed point was hand selected in a single subject's brain volume to give a good match—which was then used as the reference tract—and the seed was transferred to standard space using the inverse of the usual transformation. The resulting set of six reference tracts are illustrated in Fig. 7.3. The genu and splenium tracts were drawn from a single subject, the two cingulum bundles from another subject, and the two CSTs from two more subjects.

NT was applied in each remaining subject for each fasciculus, using a neighbourhood size of  $7 \times 7 \times 7$  voxels. Tracts segmented using the registration method and NT were inspected by eye to establish whether or not they were anatomically plausible representations of the relevant

<sup>a</sup>The corticospinal tract is the fasciculus that was segmented by seeding in the posterior limb of the internal capsule in chapter 6.





**Figure 7.3:** Reference tracts used for the ageing study, representing right (a) and left (b) cingulum bundle, right (c) and left (d) corticospinal tract, genu (e) and splenium (f). Coordinates of the original seed points in MNI space are given in each case, and native space seeds are marked with green crosses. Images courtesy of Jakub Piątkowski.

fasciculus in each case. Finally, using the field of connection probabilities associated with the selected candidate tract,  $\phi(\mathbf{x})$ , as a set of voxel weightings, tract-averaged values of  $AVF$ ,  $FA$  and  $MD$  were calculated according to

$$F = \frac{\sum_{\mathbf{x}} \phi(\mathbf{x}) f(\mathbf{x})}{\sum_{\mathbf{x}} \phi(\mathbf{x})}, \quad (7.1)$$

where  $f(\mathbf{x})$  is a scalar field encapsulating the values of  $AVF$  (and so on) at each voxel in the brain. Since we are hoping to make group contrasts more robust for comparative studies, we would hope that the variability of these measures would be smaller within this group using  $NT$  than with the registration method.

Table 7.1 shows the subjective results of examining each tract by eye to determine whether or not it represents an anatomically plausible segmentation of the relevant fasciculus. The table also shows the percentages of tracts whose segmentations were deemed better or worse using neighbourhood tractography, irrespective of whether or not the  $NT$  segmentation was actually good enough to be considered acceptable. Table 7.2 shows the coefficients of variation (cvs) for each metric, calculated using Eq. (7.1) from the single tracts selected with registration or  $NT$ . All selected tracts contributed to these values, whether or not they were found to represent acceptable segmentations.

The subjective and objective results largely corroborate one another. Coefficients of variation for each of the three  $dmRI$  metrics are generally lower using  $NT$  than they are with the registration method, except in the right corticospinal tract—which was also the only tract in which  $NT$  was judged to have worsened more tract segmentations than it improved. The cvs for  $MD$  in the left cingulum and left  $cST$  were also higher using neighbourhood tractography, but the differences in these cases were so small as to be negligible. It is clear, however, that

%	Right CB	Left CB	Right CST	Left CST	Genu	Splenium	Total
<b>RM acceptable</b>	18.5	18.5	22.2	37.0	51.9	3.7	25.3
<b>NT acceptable</b>	44.4	48.1	14.8	40.7	81.5	74.1	50.6
<b>Either acc.</b>	48.1	59.3	29.6	59.3	100.0	74.1	61.7
<b>Neither acc.</b>	51.9	40.7	70.4	40.7	0.0	25.9	38.3
<b>NT better</b>	59.3	55.6	11.1	37.0	48.1	81.5	48.8
<b>NT worse</b>	11.1	14.8	48.1	22.2	18.5	0.0	19.1

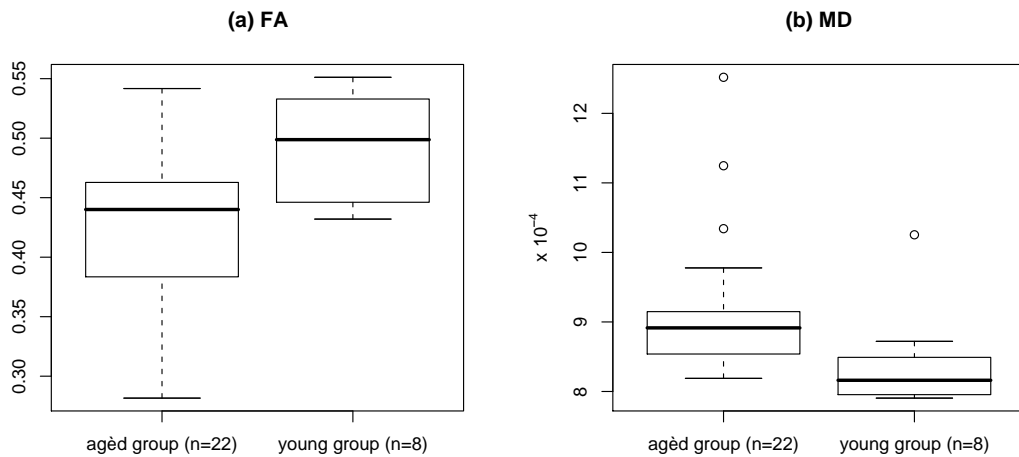
**Table 7.1:** Proportions of tracts generated by applying the registration method (RM) or neighbourhood tractography (NT) to the aged cohort which are considered “acceptable” matches, expressed as percentages. Proportions of tracts which were deemed better or worse matches after applying  $NT$  are also given. The reference tracts are included in this analysis.

Metric	Method	Right cb	Left cb	Right cst	Left cst	Genu	Splenium
AVF	RM	0.274	0.328	0.080	0.098	0.362	0.256
	NT	0.165	0.260	0.117	0.086	0.167	0.198
	difference	0.109	0.068	-0.037	0.012	0.195	0.058
FA	RM	0.242	0.298	0.075	0.095	0.282	0.159
	NT	0.152	0.229	0.113	0.083	0.141	0.136
	difference	0.090	0.069	-0.038	0.012	0.141	0.023
MD	RM	0.296	0.484	0.053	0.042	0.293	0.296
	NT	0.073	0.489	0.059	0.049	0.105	0.204
	difference	0.223	-0.005	-0.006	-0.007	0.188	0.092

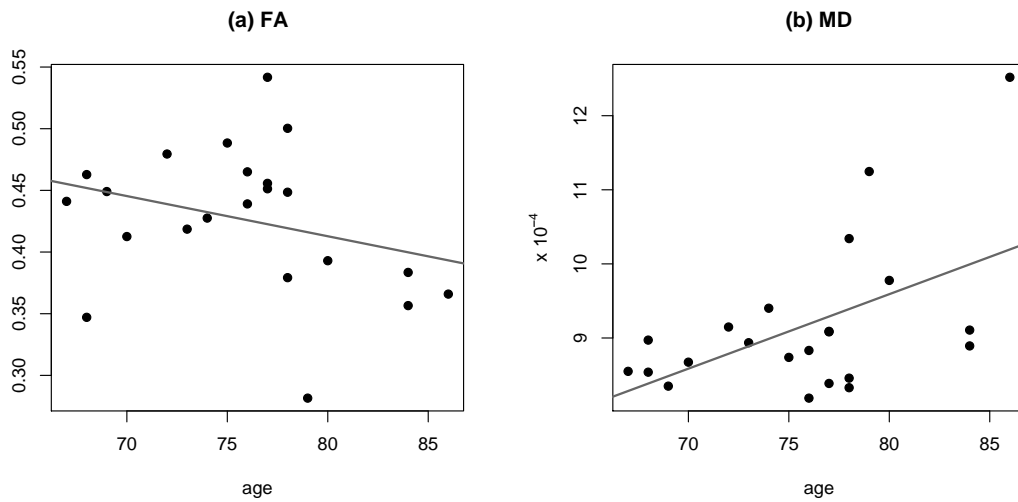
**Table 7.2:** Coefficients of variation for each metric and fasciculus, across the agèd cohort, using the registration method (RM) and neighbourhood tractography (NT). Differences are positive where the CV is greater using the registration method.

Metric	Method	Agèd mean	Young mean	<i>p</i> -value
AVF	RM	0.341	0.362	0.238
	NT	0.336	0.402	<b>0.002*</b>
FA	RM	0.429	0.447	0.359
	NT	0.427	0.492	<b>0.006*</b>
MD ( $\times 10^{-4}$ )	RM	8.99	8.52	0.275
	NT	9.16	8.42	0.053
				<i>*p</i> < 0.01

**Table 7.3:** Comparisons of the three tract metrics for the corpus callosum genu, between the agèd and young groups, using only visually acceptable segmentations. *p*-values were calculated using two-tailed *t*-tests.



**Figure 7.4:** Box-and-whisker plots of weighted mean FA and MD in the genu of the agèd and young subject groups. The thick horizontal lines represent the medians for each group.



**Figure 7.5:** Scatter plots of age against genu FA and MD within the aged subject group. Least-squares linear regression lines are shown in grey for information, but there is no significant correlation.

there was considerable variation among the fasciculi in the proportions of tracts found to be acceptable, and in the variability of tract metrics as indicated by the cv values. There are even substantial differences between bilateral pairs of tracts: twice as many tracts representing the left CST were successfully segmented using either of the two methods, for example, as for the right CST. This lack of consistency between comparable tracts across the data set may be a genuine characteristic of the data, but it is more likely that differences in reference tract quality are the main source of the effect. This is an issue that we will return to later.

## 7.2 Old versus young

We have described the effects of neighbourhood tractography in reducing the variability of diffusion metrics within a single population, but we have yet to demonstrate that this is helpful in performing group contrasts. To this end, the genu reference tract used for the study described above was used to perform NT in a group of eight young adults (mean age  $25.8 \pm 3.7$  years), using the same neighbourhood size of  $7 \times 7 \times 7$  voxels. The acquisition protocol for these subjects was described in §6.4. The three metrics of interest were calculated for the tract selected as the best match by NT, according to Eq. (7.1). These were then compared with the data from the 22 aged subjects (mean age  $75.7 \pm 5.3$  years) whose genu segmentations using NT were considered acceptable. Equivalent values were also computed using the registration method.

The results are tabulated in Table 7.3, and illustrated graphically in Fig. 7.4. We observe that the mean FA and AVF is significantly different between the groups using NT, according to a standard two-tailed *t*-test; but not with the registration method. In line with the results of Abe *et al.* (2002) and others, anisotropy is found to be higher in the younger group. The difference between MD means also approaches significance using NT, with  $p = 0.053$ . It therefore appears that NT does help with this type of contrastive analysis in specific tracts.

The box-and-whisker plots additionally give a sense of the variance in each group. We note that for both FA and MD<sup>b</sup> the interquartile range is similar for the two groups, but the full data range is considerably wider for the aged group. The greater age variation *within* the aged group is a plausible cause of these longer-tailed distributions, but further analysis did not

<sup>b</sup>The equivalent plot for AVF (not shown) closely resembles the FA plot, since the two measures are closely related.

reveal any significant age effect (see Fig. 7.5). Hence we can only conclude that there are some uncontrolled covarying factors which differ more within the older group than the younger one.

### 7.3 Improving the reference tracts

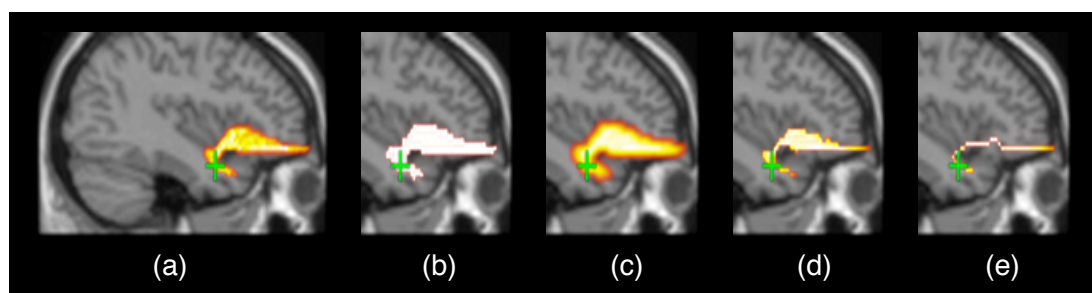
Although we have demonstrated in the previous section that a reference tract can be successfully used for segmenting tracts in multiple data sets acquired with different dMRI protocols, it must be admitted that the “hit rate”, as indicated by the proportions of tracts deemed acceptable in Table 7.1, is not especially high in the aged cohort. This will be partly due to lower data quality in this group: the reduced anisotropy will mean that orientational uncertainty is higher, and so tracking will be less reliable and more prone to diverge from the expected trajectory. Another factor is anatomical differences between subjects such as variation in ventricle size, which may tend to make the placement of the neighbourhood in native space inappropriate in some cases. A possible remedy for this is to increase the neighbourhood width. The reference tract itself, however, is an extremely important aspect of the neighbourhood tractography process; and by selecting a tract more or less arbitrarily from the data set under study we are neglecting to ensure the quality of the reference, either as a typical example of the fasciculus it represents or in terms of its optimality for the neighbourhood tractography algorithm.

In the following work, which was conducted jointly with Dr Susana Muñoz Maniega, we describe a method for defining reference tracts based on a published human white matter atlas (Mori *et al.*, 2005). These references aim to be both independent of any particular data set and carefully constructed so as to minimise ambiguity for the tract matching algorithm. This will hopefully maximise the transferability of the reference tracts, which is a major benefit of the general NT approach. In the following section we apply these references to data from the Edinburgh High Risk study (principal investigator Prof. Eve Johnstone), which involves schizophrenics and relatives considered to be at high risk of becoming schizophrenic themselves.

We begin by explaining our motivation more explicitly. The aim of a reference tract is to epitomise the topological characteristics of the fasciculus which we wish to segment in an individual brain volume. Naturally, the shape and length of the correct segmentation in any given subject’s brain will not be identical to those of any reference tract, but the tract similarity metric that we described in chapter 6 is designed to allow us to maximise the correspondence, given the constraints imposed by the data. There is, in effect, a distribution over tract topologies, from which the fasciculus of each individual is drawn. In order to maximise the effectiveness of the neighbourhood tractography method, the reference tract should represent a topology that is as close as possible to the mode of this distribution; thus ensuring that the greatest possible proportion of “correct” segmentations are considered good matches to it. A reference tract chosen from a single subject may in fact sit within the tails of the distribution—i.e. it may be an atypical outlier—even if it is appropriate for that subject, and appears to be plausible. To create a separate reference tract for each data set would also involve an undesirable and unnecessary increase in the work required to apply NT to new studies. On the other hand, atlas representations of white matter tracts are typically based on data from several subjects, and therefore give a sense of the underlying distribution.

With reference to the white matter atlas created by Mori *et al.* (2005), we manually segmented, in the MNI single subject template brain (Holmes *et al.*, 1998), the whole region corresponding to the tract of interest. We then resampled this region to correspond to the resolution of the native space in which the data for the High Risk study were acquired. Note that only a scale transformation is applied here, so this resampling process is quite subject-independent.

An example of the tract region at this stage, overlaid on an appropriately resampled image of the MNI single subject, is shown in Fig. 7.6(a). The tract in this case is the right uncinate fasciculus. This region represents all voxels in the brain through which the tract may pass, but it is considerably wider than any single tract would be. It is therefore unrepresentative, and it is also heavily suboptimal for the NT similarity algorithm, because there is no unique maximum intensity pathway through it. Our final aim is a very narrow pathway running through the centre of this region, which should be a good approximation to the mode of the



**Figure 7.6:** The steps of atlas-based reference tract generation, demonstrated on the right uncinate fasciculus. Each image is shown as a sagittal maximum intensity projection, overlaid on the slice of the MNI single subject template in-plane with the seed point. The seed is shown in green.

spatial distribution over tracts, and unambiguous for the purposes of matching. We achieve this by first binarising the image, giving all nonzero voxels the same value (b); smoothing with a Gaussian kernel with standard deviation of 2 mm, thereby encoding at each voxel the distance to edge of the region (c); and then skeletonising the result using the same principle that the TBSS technique uses for skeletonising FA maps (d). The latter skeletonisation process works by finding local maxima in image intensity (cf. §6.1). What remains is a “core” of the original region, from which a reduced tract is calculated (cf. §6.3.1), producing the final reference tract (e), which has single voxel thickness along its length and is therefore unambiguous in orientation at each step of the matching algorithm. Seed points for these reference tracts are placed to avoid regions where fibres are expected to cross, or where contaminating tract orientation information might otherwise be expected to be present.

## 7.4 A schizophrenia study

Evidence from functional imaging has led to the suggestion that schizophrenia may be a disconnection syndrome, in which interaction between frontal and temporal regions is particularly abnormal (Friston & Frith, 1995). As a result, there is a considerable literature of white matter studies in schizophrenia, and dMRI methods are now commonly applied as part of them. Voxel-based analyses have provided evidence of dMRI-visible changes in the uncinate and arcuate fasciculi (Burns *et al.*, 2003) and cingulum bundle (Kubicki *et al.*, 2003), amongst other regions. Park *et al.* (2004) also demonstrated consistent hemispheric asymmetries in the anisotropy of a number of white matter structures, in both healthy and schizophrenic subjects.

The use of tractography in studies of schizophrenia has so far been limited. Kanaan *et al.* (2006) and Price *et al.* (2007) both use tractography methods to demonstrate reduced corpus callosum FA in schizophrenics, while Jones *et al.* (2006) examine a number of tracts but find a significant difference in FA only in the left superior longitudinal fasciculus.<sup>c</sup>

In preparation for this study, reference tracts for the two cingulum bundles, arcuate fasciculi (AF), uncinate fasciculi (UF) and anterior thalamic radiations (ATR) were created as described above. The latter fasciculus connects prefrontal cortex to the thalamus through the anterior limb of the internal capsule—its pertinence is due to evidence of reduced anterior thalamic grey matter density in schizophrenia (McIntosh *et al.*, 2004), which might be linked to a breakdown of connectivity between this part of thalamus and cortex.

27 schizophrenic patients (mean age  $36.5 \pm 9.2$  years), 20 healthy subjects at an enhanced risk of becoming schizophrenic due to having relatives with the disorder (mean age  $30.2 \pm 2.9$  years) and 50 healthy controls (mean age  $35.3 \pm 10.9$  years) underwent the dMRI protocol described in §6.4. These data were acquired by Dr Dominic Marjoram and Dr Andrew McIntosh. For each of the eight tracts of interest, neighbourhood tractography was applied to each subject, using a

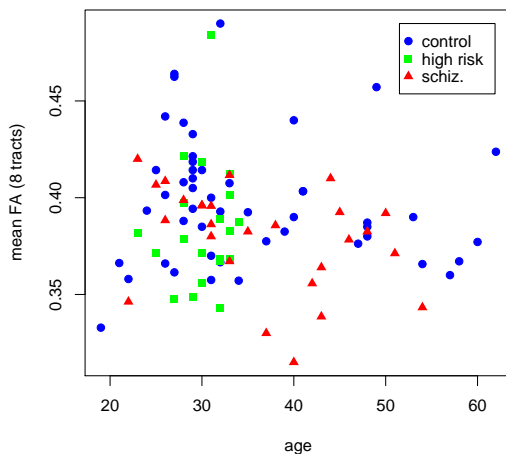
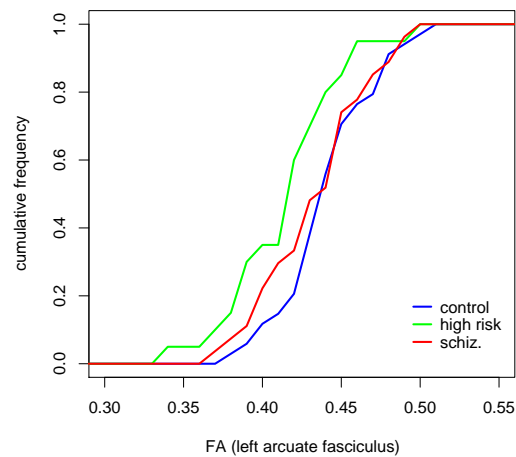
<sup>c</sup>The superior longitudinal fasciculus and arcuate fasciculus are closely related structures, and the names are often used interchangeably; although recent work suggests that they should not be considered identical (Makris *et al.*, 2005). It is not clear what definition of the fasciculus is being used by Jones *et al.* in this context.

Tract	% acceptable	Control mean FA	HR mean FA	Schiz. mean FA	<i>p</i> -value
Right CB	73.2	0.407	0.384	0.406	0.407
Left CB	83.5	0.430	0.413	0.390	0.111
Right AF	70.1	0.418	0.423	0.406	0.395
Left AF	83.5	0.444	0.419	0.437	<b>0.028*</b>
Right UF	94.8	0.364	0.354	0.347	0.258
Left UF	92.8	0.380	0.357	0.357	0.059
Right ATR	71.1	0.355	0.347	0.342	0.719
Left ATR	77.3	0.389	0.367	0.362	0.096

\* $p < 0.05$ 

**Table 7.4:** Bilateral results based on weighted FA values calculated in the cingulum bundles (CB), arcuate fasciculi (AF), uncinate fasciculi (UF) and anterior thalamic radiations (ATR). Group means were calculated for control, high risk (HR) and schizophrenic subjects and compared using a one way ANOVA. *p*-values given are derived from a standard *F*-test.

**Figure 7.7:** Cumulative frequency plots of weighted mean FA in the left arcuate fasciculus of each group. Bonferroni corrected *t*-tests found a significant difference between high risk subjects and controls only.



**Figure 7.8:** Scatter plot of age against mean FA across all of the tracts used in this study.

neighbourhood width of  $7 \times 7 \times 7$  voxels as before. The proportions of visually plausible tracts were recorded in each case, and for these acceptable segmentations, a weighted mean  $FA$  value was calculated as per Eq. (7.1). For each tract, a one way analysis of variance (ANOVA) was applied to establish whether there was any effect of group membership on anisotropy. We also examined the relationship between age and anisotropy, averaged over all tracts, for each group individually.

ANOVA results are given in Table 7.4. We observe that mean  $FA$  in controls is higher than the other two groups in seven of the eight tracts—the right arcuate fasciculus is the only exception—but there is a significant effect of group membership only in the left arcuate. This result is consistent with the findings of Jones *et al.* (2006), but post-hoc  $t$ -tests applied to these data showed that the significant difference was between controls and the high risk group, with a Bonferroni corrected  $p$ -value of 0.036. The mean in the schizophrenic population was considerably higher than the high risk mean for this tract. The lack of significance between schizophrenics and controls might be related to greater variance in the former population, but a cumulative frequency plot (Fig. 7.7) does not bear this hypothesis out. The general steepness of the curve—which hints at the spread of the data—is similar between the control and schizophrenic populations. There is perhaps a tendency for  $FA$  values below the group median to be lower in schizophrenics than controls, but in general the two curves are genuinely very similar. The high risk curve, by contrast, is consistently shifted towards lower  $FA$  values.

We additionally note, in common with Park *et al.* (2004), that there is a noticeable lateralisation effect in mean  $FA$ s, which are invariably higher, on average, in the left hemispheric versions of each tract.

Fig. 7.8 shows a scatter plot of age against the average  $FA$  across all tracts. Our set of eight tracts was not identical to those used by Jones *et al.* (2006), but we nevertheless failed to find evidence of the general age effect described in that study. Of the three groups, only the schizophrenics yielded a statistically significant relationship (Spearman's  $\rho = -0.43$ ,  $p = 0.024$ ), but the correlation was negative in this case, not positive as in Fig. 7.1.

There are any number of reasons that might help to explain why relationships between clinical status and tract  $FA$  were not more numerous. There may be genuinely little effect on white matter;  $FA$  may not be sensitive to the kinds of physiological abnormality associated with schizophrenia, or only inconsistently so; or the effect may be so small that it is masked by noise. Jones *et al.* (2006) suggest that the age of onset of schizophrenic symptoms may be a relevant covarying factor to include in a more complex analysis. Since there was no difference between controls and schizophrenics in the left arcuate fasciculus, it is difficult to interpret the finding of difference between the control and high risk groups. Because the latter was not very strongly significant, it may be simply coincidental.

Despite a paucity of clinical findings under the relatively simple analysis that we have applied here, the considerably higher acceptance rates—reaching up to 95%—for tracts segmented using atlas-based reference tracts are encouraging. Of course, it would be necessary to use these tracts in the aged cohort in order to make a direct comparison between the two types of reference—the subjects involved in this study are, after all, noticeably younger. However, even if the improvement is robust, it is not yet large enough to allow us to dispense with manual checking of the selected tracts; and in small subject groups the rejection rate may still be considered unacceptably high. The limitations of the similarity measure discussed in §6.6 remain in any case. There is certainly room for improvement in the neighbourhood tractography method itself.

## 7.5 Conclusions

Despite some evidence of greater variability among the older volunteer population that we studied in the first half of this chapter, compared to the younger volunteer group, we have found that a reference tract drawn from one group can be used to successfully guide the selection of candidate tracts in the other. However, the proportions of tracts successfully segmented using a reference drawn from the data set, as estimated by a human observer, was somewhat lower than might be hoped. A marked improvement was found using reference

tracts based on a white matter atlas—although different raters were involved in these two studies, so some of the difference may be attributable to inconsistency in acceptance criteria.

The discussion of intersubject distributions over tracts in this chapter raises the possibility of using a formal probabilistic model to represent this variability. With proof of concept for neighbourhood tractography in place, refinement of the method is our next priority.



# Model-based tract matching

THE FORMULATION of neighbourhood tractography described in chapter 6 has several intrinsic limitations, and it is also relatively inscrutable because of its essentially heuristic formulation. In this chapter we describe an attempt to formalise the principle of neighbourhood tractography into a probabilistic model, and use machine learning methods to find matching tracts from a set of candidate tracts.

We move to representing tracts in terms of single lines, and describe explicit probability distributions to encapsulate the variability in shape and length across subjects. The parameters of the resulting model are fitted using maximum likelihood from a number of hand-picked training tracts, and then used to select matching tracts in separate test cases. We later go on to describe a similar but unsupervised method, which negates the need for separate training data. These approaches are found to overcome the main limitations of the heuristic method.

## 8.1 B-splines

This chapter will make use of *B-splines*, which are a type of parametric curve commonly used in computer graphics, and a generalisation of the *Bézier curve* (Böhm *et al.*, 1984; de Boor, 1978). Both B-splines and Bézier curves are linear combinations of polynomial basis functions, whose general form can be expressed as

$$\mathbf{r}(t) = \sum_{i=0}^p \mathbf{P}_i B_{i,n}(t), \quad (8.1)$$

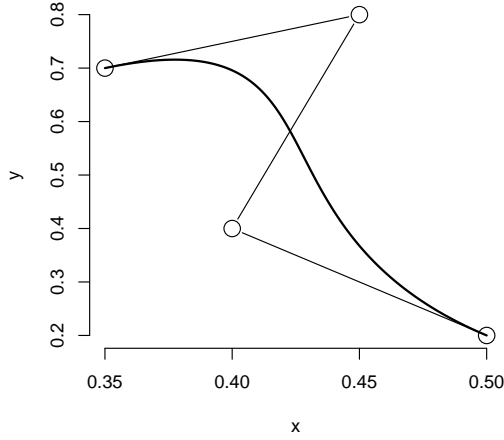
where  $B_{i,n}$  are the basis functions of degree  $n$  and the coefficients,  $\mathbf{P}_i$ , for  $i \in \{0..p\}$ , are called control points. The parameter  $t$  is conventionally taken to be in the normalised interval  $[0, 1]$ . The curve can be defined in as many dimensions as are required, by providing control point vectors of the appropriate dimensionality.

In the relatively simple case of a Bézier curve, the basis functions are the family of polynomials given by

$$B_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i},$$

where  $\binom{n}{i}$  is a binomial coefficient,  $n$  is the degree of the polynomial, and  $i \in \{0..n\}$ . An example of a two-dimensional curve built up from this basis, with  $n = 3$ , is given in Fig. 8.1. Note that the first and last control points coincide with the line, while the others guide its direction and curvature. Any Bézier curve with degree  $n$  has exactly  $n + 1$  control points; so  $p = n$  in Eq. (8.1).

There are a number of advantages in representing a smooth curve in this way. Firstly, given any particular choice for the degree of the basis functions, the control points are sufficient to specify the path of the curve exactly—a far more efficient and scalable representation of the curve than a series of very short straight lines connected together (the piecewise linear representation). Secondly, the curve can be translated or rotated by applying the required transformation to the control points.



**Figure 8.1:** Two-dimensional cubic Bézier curve (thick line) with control points indicated with open circles.

B-splines follow a similar principle, but introduce the additional notion of a *knot*. Any given B-spline is associated with a sequence of knot points,  $(t_j)$ , with  $j \in \{0..m\}$ . This sequence is constrained to be nondecreasing, so that  $t_j \leq t_{j+1}$  for all appropriate values of  $j$ . The basis functions are defined iteratively, with the base case

$$B_{j,0}(t) = \begin{cases} 1 & \text{if } t_j \leq t < t_{j+1} \\ 0 & \text{otherwise.} \end{cases} \quad (8.2)$$

The recursive definition for all basis functions of higher degree is then given by

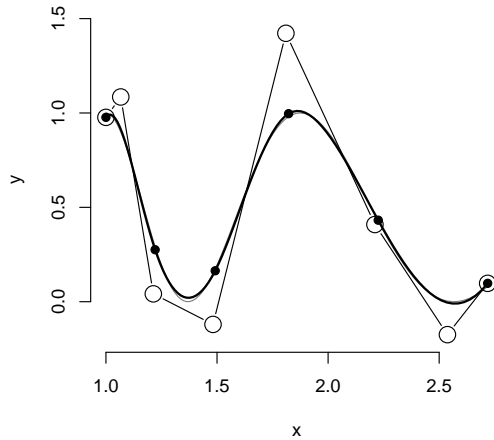
$$B_{j,n}(t) = \frac{t - t_j}{t_{j+n} - t_j} B_{j,n-1}(t) + \frac{t_{j+n+1} - t}{t_{j+n+1} - t_{j+1}} B_{j+1,n-1}(t). \quad (8.3)$$

Unfortunately it is far from obvious exactly what form the basis functions take, given this method of defining them; a problem that is exacerbated by the fact that the functions themselves depend on the knot locations,  $t_j$ , in the spline. So rather than explicitly expanding the functions for a particular case, we note their most important properties below.

- The function  $B_{j,n}(t)$  is defined over the interval  $[t_j, t_{j+n+1})$ . It is zero everywhere else.
- $B_{j,n}(t)$  is made up of  $n + 1$  polynomials of degree  $n$ , which meet at the knot points.
- The basis functions of any given degree always sum to unity:  $\sum_j B_{j,n}(t) = 1$ .
- If  $t_j$  is always strictly less than  $t_{j+1}$ , then  $B_{j,n}(t)$  is  $n - 1$  times differentiable at knot points. This means that a linear B-spline is continuous only in value at knot points, while a quadratic B-spline is also continuous in gradient, and so on.

The first of these represents a notable difference between a B-spline and a Bézier curve: the shape of the latter is affected everywhere by all control points, whereas the B-spline's control points affect the curve only locally. To clarify, then: a knot marks a boundary between basis functions, whilst a control point guides the shape of the spline.

If two or more consecutive knots fall on exactly the same value of  $t$ , the last of the properties described above is no longer true. Rather, if there are  $k$  copies of a particular knot value, then the curve is differentiable only  $n - k$  times at that point. The knot sequence is often arranged such that the first  $n + 1$  knots are 0, and the last  $n + 1$  knots are 1. This makes the curve not only nondifferentiable, but also discontinuous, at its extrema; and as a result the first and last control points directly define the curve's start and end points, as for the Bézier case. The remaining knot points are known as *internal knots*. A B-spline with no internal knots is a Bézier curve.



**Figure 8.2:** Cubic B-spline approximation to the 2-D parametric function  $(e^t, \cos^2 5t)$  over the interval  $[0,1]$ , using four internal knots. The thick black line is the spline curve, filled circles represent knot points and open circles control points. The true function is shown in grey, but it is barely visible because the B-spline approximates it very closely.

An example of a B-spline is shown in Fig. 8.2. In this case the spline has four internal knots and four repeated knots at each end, for a total of twelve. There are  $m - n$  cubic basis functions—i.e. eight in this case—and therefore eight control points can be seen in the figure, two of which coincide with the repeated knots at the curve extrema. It can be seen that the curve approximates two nonpolynomial functions to a high degree of accuracy, and the approximation could be improved still further by adding additional knots, each of which will increase the number of control points and thus the degrees of freedom of the parameterisation.

If the knot vector is known then the control points are sufficient to reconstruct the complete spline curve. For a *uniform* B-spline, where the internal knot points are equally spaced across the range of  $t$  values, the knot vector is highly constrained and even less information is therefore required to recover the curve.

## 8.2 Tract representation revisited

It is desirable, as we discussed in §6.3, that the tract representation chosen for the purpose of matching be as independent as possible of the fibre tracking algorithm used to generate the tract. In our original formulation of neighbourhood tractography, we chose to work with tracts represented as a scalar field over the native space of each subject, with an associated seed point—a form amenable to both probabilistic and deterministic algorithms. However, two of the greatest limitations of the matching algorithm outlined there arise from this choice: the difficulty of correcting for gross rotation and scaling differences between subjects, and the risk of premature termination due to local directional uncertainty in the reference or candidate tract. The process of calculating a reduced tract is also susceptible to this latter problem.

In the following work, we use a B-spline tract representation instead. This choice necessitates some loss of information when the original tract was made up of many sample streamlines, but this loss is only for the purpose of matching, so it need not entail major difficulties provided that sufficient information remains to meaningfully compare the shape and length of a candidate and a reference tract.

To recap: whether a reconstructed tract consists of a single line running through a seed point, or a number of sample streamlines with the seed point in common, the process for generating streamlines is typically to choose a local tract orientation—starting at the seed point—move a short distance in the corresponding direction, and repeat until some termination criterion is met. This process has to be performed twice to reconstruct the complete streamline, since all dMRI-derived tract orientation information is directionally nonspecific. As a result, each streamline can be conceptually split at the seed point into two sets of points, representing what we will refer to as the “left” and “right” substreamlines. The streamline can therefore be said

to have a “left length”,  $N_1$ —the number of points on its left side, excluding the seed point itself—and a “right length”,  $N_2$ . Note that the names left and right are used for convenience only, and have no strict significance.

In order to be able to model single streamlines and distributions of probabilistic streamlines in the same way, we must first find a single line, in the latter case, which epitomises the shape of the whole set of lines. We do this by calculating a median streamline whose left and right lengths,  $\tilde{N}_1$  and  $\tilde{N}_2$ , are the  $\xi$ -quantiles of the individual streamline lengths, where  $\xi$  is a parameter to be chosen. (For  $\xi = 0.9$ , for example, distal spatial information would be discarded from the longest 10% of streamlines.) Then, beginning at the seed point and moving outwards in each direction in turn, the  $x$ ,  $y$  and  $z$  components of the median point location are calculated at each step from all untruncated streamlines. The resultant set of median points is a single line tract representation  $r = (\mathbf{x}_i)$ , where  $i \in \{-\tilde{N}_1, -\tilde{N}_1 + 1, \dots, \tilde{N}_2 - 1, \tilde{N}_2\}$  and the point  $\mathbf{x}_0$  is the seed point. (Alternatively, a single streamline could be extracted from the data set by minimising any of the distance metrics mentioned in §6.2 within the set.)

Unlike in the individual streamlines, where the distance between successive points is fixed, the median line, as a composite streamline, is not in general made up of equally spaced points. In fact, since the number of streamlines drops as one moves away from the seed point, and the median location is calculated from only untruncated streamlines, it may occasionally move a large distance in a single step. Nevertheless, the real world length of this piecewise linear median line can, of course, be calculated by summing the actual point spacings.

Finally, the path of the median line is represented in terms of a three-dimensional cubic B-spline curve, parameterised by the distance along the line,  $t$ . For any uniform cubic B-spline with  $m + 1$  knots in total, there are  $\kappa = m - 7$  equally spaced internal knots; and in this case they are arranged so that one of them falls on the seed point. The final tract parameterisation then becomes

$$\mathbf{r}(t) = \sum_{j=0}^{m-4} \mathbf{P}_j B_{j,3}(t), \quad (8.4)$$

a particular case of Eq. (8.1), where  $B_{j,3}$  are the cubic B-spline basis functions.

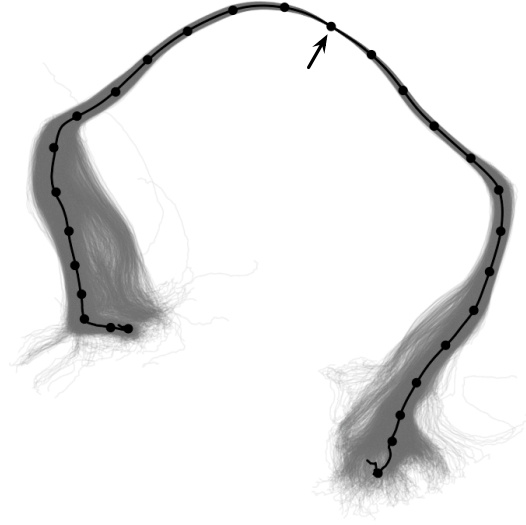
The free parameter,  $m$ , is not chosen directly. Instead, the control point coefficients are calculated for the reference tract data using a model with one internal knot (i.e.  $\kappa = 1$ ,  $m = 8$ ), and the residuals,  $\rho_i$ , at each point,  $i$ , on the median line are used to calculate the residual standard error, according to

$$\mathbf{E}_\kappa = \sqrt{\frac{\sum_i \rho_i^2}{\tilde{N}_1 + \tilde{N}_2 - \kappa - 3}}. \quad (8.5)$$

(The denominator of Eq. (8.5) represents the residual degrees of freedom, which is affected by the number of points on the median line and the number of internal knots.) The number of knots is then incremented and the residual standard error recalculated until the mean of the three components of  $\mathbf{E}_\kappa$  is less than some threshold value,  $\eta$ . The knot separation distance for this fit is then fixed for each candidate tract, so the number of knot points in each case depends on—and is uniquely determined by—the length of each median line.

Fig. 8.3 demonstrates the process described above. A set of 5000 probabilistic streamlines is shown in grey: these represent all of the information about the connectivity distribution provided by the tractography algorithm for a single seed point. The black line represents the median, and the black filled circles represent the B-spline knot points in the final tract parameterisation. Note that, although we favour methods that produce a distribution of streamlines due to the greater amount of information they provide about spatial uncertainty, if a tractography algorithm had been used that generates only a single streamline for each seed point, then calculating the median line would be unnecessary, but the B-spline parameterisation would still be valid. This parameterisation is used in order to reduce the dimensionality of the data and emphasise topological tract features at a scale that is not determined by voxel dimensions.

**Figure 8.3:** Graphical representation of a full set of probabilistic streamlines representing the corpus callosum splenium (grey, shown at 10% opacity), the median line and B-spline knot points (black), here projected into a plane normal to the superior–inferior (z) axis.  $\xi = 0.99$ . The seed point is indicated with an arrow.



### 8.3 Comparing spline tracts

With the reference and candidate tracts represented as B-splines, we can now define a model for the topological relationships between them. We consider a finite set of candidate tracts, among which there is assumed to be a single tract that best matches the reference tract, which has been chosen in advance. We introduce a variable,  $\mu$ , which can take any value in  $\{1..N\}$ , where  $N$  is the number of candidate tracts in the set, to indicate that the corresponding tract is the best match. Given a set of data,  $D$ , describing a group of candidate tracts, we wish to establish a model for the distribution  $P(\mu|D)$ ; and hence to find the most likely value of  $\mu$ .

For a tract,  $i$ , which has  $L_1$  internal knot points on its left side and  $L_2$  internal knots on its right side—excluding the seed point in each case—we consider the vectors that link successive knots together such that they are always directed away from the seed point. We denote these vectors  $\mathbf{v}_u^i$ , where  $u$  indexes over knot points such that it is negative on the left side of the tract and positive on the right side. The cosine of the angle between a contiguous pair of these vectors is given by

$$c_u^i = \cos \theta_u^i = \frac{\mathbf{v}_u^i \cdot \mathbf{w}_u^i}{\|\mathbf{v}_u^i\| \|\mathbf{w}_u^i\|} \quad (8.6)$$

where  $\|\cdot\|$  is the usual Euclidean norm and

$$\mathbf{w}_u^i = \begin{cases} \mathbf{v}_{u+1}^i & \text{if } u < -1 \\ -\mathbf{v}_{-u}^i & \text{if } u = \pm 1 \\ \mathbf{v}_{u-1}^i & \text{if } u > 1. \end{cases}$$

These *continuity* angles give an indication of the local curvature of the tract. By introducing the notation  $\mathbf{v}_u^*$  for the  $u$ th vector in the reference tract, we can describe another cosine value,

$$s_u^i = \cos \phi_u^i = \frac{\mathbf{v}_u^i \cdot \mathbf{v}_u^*}{\|\mathbf{v}_u^i\| \|\mathbf{v}_u^*\|}, \quad (8.7)$$

which indicates the local directional *similarity* between the reference and candidate tracts.

Fig. 8.4 illustrates, in two dimensions, the continuity angles,  $\theta_u$ , and the similarity angles,  $\phi_u$ . The cosine function is *a priori* uniform in the sense that the distribution of cosines between pairs of vectors generated from a spherically symmetric distribution is uniform; and it is therefore convenient to model the continuity and similarity cosines, as described by Eqs (8.6) and (8.7), rather than directly modelling the angles themselves.

The tract data that are relevant to our matching model are its continuity and similarity cosines and its left and right lengths:  $\mathbf{d}^i = (L_1^i, L_2^i, \mathbf{c}^i, \mathbf{s}^i)$ , where  $\mathbf{c}^i = (c_u^i)$  and  $\mathbf{s}^i = (s_u^i)$ . The full

data set,  $D$ , then consists of all the  $\mathbf{d}^i$  plus the left and right lengths of the reference tract,  $L_1^*$  and  $L_2^*$ . The principle of the model is that in regions where there is directionality information available from the reference tract, that information should provide the best predictor for the direction of a matching candidate tract. If the candidate tract is longer than the reference tract, however, then in the region beyond the end of the reference, the only predictor of the tract's direction at any given step is its direction at the previous step. Hence, the full matching model is given by

$$P(\mu = i | D) \propto P(L_1^i | L_1^*) P(L_2^i | L_2^*) \prod_{u=1}^{\check{L}_1^i} P(s_{-u}^i) \prod_{u=1}^{\check{L}_2^i} P(s_u^i) \prod_{u=\check{L}_1^i+1}^{L_1^i} P(c_{-u}^i) \prod_{u=\check{L}_2^i+1}^{L_2^i} P(c_u^i), \quad (8.8)$$

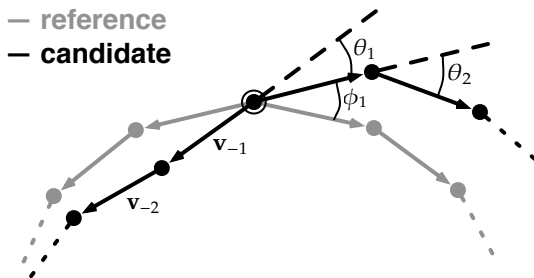
where  $\check{L}_1^i = \min\{L_1^i, L_1^*\}$ , and equivalently for  $\check{L}_2^i$ . The inclusion of the continuity cosine distributions expresses a preference for candidates that are not atypical in their curvature in regions unconstrained by the reference tract; it thus provides some assurance of “tract quality”. It is implicitly assumed here that all unmatched tracts are equiprobable. The constant of proportionality in Eq. (8.8) is given by normalising over all values of  $i$ .

There are some constraints that can be applied to this model in order to reduce the number of parameters that need to be estimated. To this end, we assume that the curvature properties of tracts do not vary along their length, implying that all continuity cosines are drawn from a single distribution. We cannot, however, assume the same for the similarity cosines: Fig. 8.3 demonstrates that there is generally far more spatial uncertainty—as shown by the spread of the streamline set—near the ends of tracts than there is near the middle, so considerable local deviation from the reference tract can be expected near the ends of even well-matched candidate tracts. Hence, we make the weaker assumption that there is no inherent difference between the left and right sides of the tract, with distributions over similarity cosines varying only with distance from the seed point. That is,

$$\begin{aligned} P(c_u^i) &= P(c_v^i) = P(c) & \forall u, v, i \\ P(s_u^i) &= P(s_{-u}^i) = P(s_u) & \forall u > 0, i. \end{aligned} \quad (8.9)$$

We must finally give specific forms for the distributions in Eq. (8.8). The length distributions are modelled as regularised multinomial distributions, subject to a maximum length cutoff. Fitting such a model from a data set using maximum likelihood is almost trivial: one simply counts the number of times each length value occurs in the data set, adds a small constant value to each count to regularise the distribution, and then normalises. The regularisation ensures that the matching probability is not zero for a tract whose exact left and right lengths were not in the training data set, which would be a strong and unjustified imposition.

The cosine distributions are less straightforward. If there were no relationship between the reference and candidate tracts then the similarity cosines would be approximately uniformly distributed, as we discussed above. However, if smaller deviations from the reference tract are assumed to be far more common than larger ones, as we expect for matching tracts, then the distribution over cosines will be strongly biased towards the higher end. A standard distribution that is able to represent this kind of relationship over a fixed interval is the beta



**Figure 8.4:** Illustration of the different angles relevant to our model. Filled circles here represent successive knot points in the reference and candidate tracts. The ringed knot is the seed point, which is common to the two tracts.

distribution, whose general p.d.f. is given by

$$P(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad \text{for } 0 \leq x \leq 1,$$

where  $\Gamma(\cdot)$  is the gamma function, and  $\alpha$  and  $\beta$  are parameters. However, since small angles are always assumed to be the most common, we can fix  $\beta = 1$ . We also need to rescale the cosine values into the interval  $[0, 1]$  over which the distribution is defined. Finally, in order to ensure that the model does not grossly underestimate matching probabilities when larger angles occur, we add a uniform component to regularise the distribution, resulting in the mixture model

$$P(x) = \frac{1}{2} \left( \varepsilon + (1 - \varepsilon) \alpha \left( \frac{x+1}{2} \right)^{\alpha-1} \right) \quad (8.10)$$

for both the continuity and similarity cosines. This distribution becomes uniform when either  $\alpha = 1$  or  $\varepsilon = 1$ , and is strongly biased for small  $\varepsilon$  and large  $\alpha$ .

To find maximum likelihood estimates for  $\alpha$  and  $\varepsilon$  given some data vector of rescaled cosine values,  $\mathbf{x}$ , we use a simple Expectation–Maximisation algorithm. Associated with each data value,  $x_j$ , is a latent variable,  $\zeta_j$ , indicating whether the value came from the uniform distribution ( $\zeta_j = 0$ ) or the beta distribution ( $\zeta_j = 1$ ). Given some starting estimates for the distribution parameters,  $\hat{\alpha}$  and  $\hat{\varepsilon}$ , the E-step of the algorithm calculates

$$P(\zeta_j = 0|x_j) = \frac{\hat{\varepsilon}}{\hat{\varepsilon} + (1 - \hat{\varepsilon}) \hat{\alpha} x_j^{\hat{\alpha}-1}}$$

and

$$P(\zeta_j = 1|x_j) = \frac{(1 - \hat{\varepsilon}) \hat{\alpha} x_j^{\hat{\alpha}-1}}{\hat{\varepsilon} + (1 - \hat{\varepsilon}) \hat{\alpha} x_j^{\hat{\alpha}-1}}$$

for each value of  $j$ . The M-step then updates the parameter estimates according to

$$\hat{\alpha} = \frac{-\sum_j P(\zeta_j = 1|x_j)}{\sum_j P(\zeta_j = 1|x_j) \ln x_j}$$

and

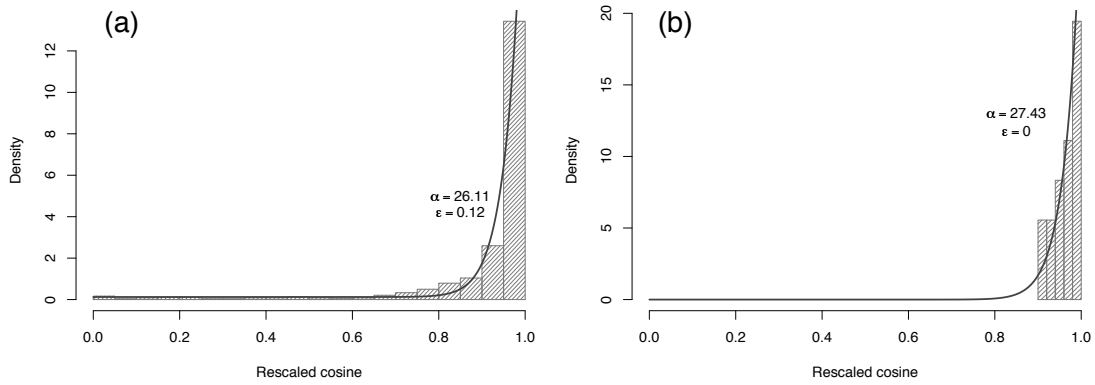
$$\hat{\varepsilon} = \frac{P(\zeta_j = 0|x_j)}{P(\zeta_j = 0|x_j) + P(\zeta_j = 1|x_j)},$$

and the algorithm repeats until convergence.

## 8.4 Training and using the model

The data used for testing this approach were those acquired for the original neighbourhood tractography experiments, taken from 14 dMRI scans of 6 individual subjects. The MRI acquisition protocol can be found in §6.4. Preprocessing to extract the brain and correct for eddy current induced distortions was performed as described there.

For the purposes of this study, the white matter structures of interest were the corpus callosum splenium and corticospinal tract. All tracts were generated using the BEDPOST/ProbTrack algorithm (Behrens *et al.*, 2003b) with its default parameters. The result was a set of 5000 probabilistic streamlines for each tract, with a fixed separation distance of 0.5 mm between successive points. Median lines were then calculated using  $\xi = 0.99$ , and transformed into the space of the reference tract, using the FLIRT registration algorithm (Jenkinson & Smith, 2001) to register together  $T_2$ -weighted ( $b = 0$ ) volumes from each scan. Using a residual error threshold,  $\eta$ , of 0.1 mm, the B-spline parameterisation was calculated for the splenium reference tract, and all candidate tract splines were fitted using the resulting knot separation distance of 6.1 mm. If any two successive median line points were more than this distance apart, the median line



**Figure 8.5:** Histograms of rescaled (a) continuity cosines ( $n = 962$ ) and (b) similarity cosines ( $u = 7$ ,  $n = 18$ ) from the splenium training data. The appropriate density functions from the model are overlaid.

was truncated to avoid creating multiple knots, which would result in discontinuities in the spline.

In addition to the reference, nine other splenium tracts were chosen by hand from different brain volumes to form a training set of matching tracts, and the parameters of the model pertaining to the length and similarity cosine distributions were fitted using maximum likelihood as described above. Specifically, three splenium tracts were taken from subject 1, two from subject 2, two from subject 3, and one each from subjects 4 and 5. The reference tract was taken from a third scan of subject 2. No more than one training tract was taken from any given scan. The continuity cosine distribution,  $P(c)$ , was fitted from 50 tracts generated by seeding randomly in a single brain volume, subject to an anisotropy threshold used to ensure that each seed point was in white matter. This policy is appropriate given the assumption that the continuity properties of all tracts are broadly similar, and it has the significant advantage of increasing the quantity of training data available.

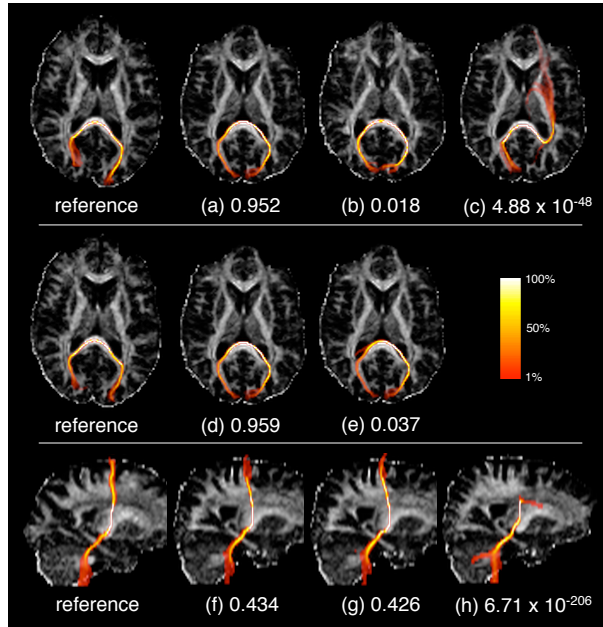
Fig. 8.5 shows histograms of the cosine distributions,  $P(c)$  and  $P(s_u)$ —the latter for a sample value of  $u$ . In (a), there are data from the full domain of (rescaled) cosine values, and the final estimate for  $\epsilon$  reflects this. In (b), however, there are no cosine data below 0.9, and so the  $\epsilon$  parameter has shrunk to zero. In fact, all of the similarity cosine distributions had  $\epsilon = 0$ , although the  $\alpha$  parameter—which affects the steepness of the right hand sides of the distributions—varied considerably, being 112.6 for  $u = 1$  and only 6.1 for  $u = 14$ , the largest value of  $u$  for which a distribution was defined.

The whole process was applied in the same manner for the corticospinal tract, using an appropriate reference. The model parameters were retrained for this case, using a training set of five tracts.

Having used the training data to learn its parameters, the model described by Eq. (8.8) represents a way of assessing a set of novel tracts for their respective similarities to the reference tract. In order to create such a set, the seed point used to generate the reference tract was transferred to a new brain volume, from which no training data had been taken. Tractography was then performed for all points within a  $7 \times 7 \times 7$  voxel region centred at this location, subject to an anisotropy threshold, and each candidate seed point was processed as follows.

1. Run the tractography algorithm and recover a set of probabilistic streamlines.
2. Calculate the median line and transform it into the space of the reference tract as described above.
3. Using the fixed knot spacing chosen, fit a cubic B-spline along the median line.
4. Calculate continuity and similarity angles for the interknot vectors, as depicted in Fig. 8.4.





**Figure 8.6:** The two most likely matches to the original (top row) and the alternative (middle row) splenium reference tract, shown in axial projection with their associated matching probabilities. The tract generated from the neighbourhood centre point is shown with its matching probability (c), for comparison. Results for the corticospinal tract, in sagittal projection, are shown in the bottom row. It should be remembered that tracts (a–h) are taken from different subjects to the reference tracts. Colours represent the proportion of probabilistic streamlines passing through each voxel, as indicated by the colour bar.

5. Evaluate the right hand side of Eq. (8.8) using the length and angle distributions fitted from the training data.

This allows us to select the “best” seed point *a posteriori* by finding the starting location which generates the best matching tract.

In order to test the robustness of the method to small differences in the reference tract, the corpus callosum reference was substituted for its equivalent taken from a different scan of the same subject (see Fig. 8.6). These two tracts do, of course, represent the same physical fasciculus, imaged in two consecutive scans. The model parameters were then recalculated for this alternative reference tract, and the experiment was repeated.

Fig. 8.6 shows the results of applying the model to tract—and hence seed point—selection. In this figure, all tracts are shown as maximum intensity projections; splenium tracts in a plane normal to the superior–inferior (*z*) axis, and corticospinal tracts normal to the left–right (*x*) axis. These perspectives are used because they show the two axes of greatest spatial variation and highlight the most common gross reconstruction inconsistencies in each case. Each tract is shown colour-coded according to the proportion of probabilistic streamlines that pass through each image voxel, thresholded at the 1% level. (This threshold is approximately equivalent to the use of  $\xi = 0.99$  above in calculating the median line.) The underlying greyscale image in each case is the slice of the anisotropy map in-plane with the seed point.

According to the model, tracts (a) and (b) are the two most likely matches to the reference tract adjacent to them. The point at the centre of the seeding neighbourhood generated tract (c), which is visually far less similar to the reference tract. Its matching probability is commensurately smaller, by many orders of magnitude, than those for (a) and (b). The candidate set contained 220 tracts in total, after thresholding on anisotropy.

For comparison, tracts (d) and (e) are the two best-matching tracts from the same neighbourhood, using the alternative reference tract. In this case the model parameters were relearned, but the knot separation distance under this very similar reference tract was only slightly smaller than the old one, at 6.0 mm. Tracts (a) and (d) are in fact the same tract, so the most likely match is the same with both reference tracts.

Similarly, tracts (f) and (g) are much better matches to the corticospinal reference tract than the tract generated from the centre seed, (h). Once again the matching probabilities reflect this.

Since there is no normalisation or standardisation of matching probabilities between different sets of candidate tracts, these values are not directly comparable between data sets or reference tracts. They simply represent the probability of each candidate tract matching the given reference *relative* to the other candidates. There is no guarantee that the most likely match

is a good match in any absolute sense. In order to provide an indication of absolute goodness-of-match, the log-ratio between the matching likelihood—the right hand side of Eq. (8.8)—of the best match and the matching likelihood of the reference tract to itself was calculated.

Fig. 8.7 shows the results of calculating log-ratios using the original reference tract for the splenium. The more negative this log-ratio, the less good a fit is compared to the “benchmark” of the reference tract itself.

## 8.5 Advantages and limitations

Compared to the simpler methods of placing single seed points by hand or using image-registration-based transformation, our method offers advantages with respect to consistency and reproducibility. As with all neighbourhood tractography methods, reference tracts can be directly transferred between studies with minimal modification; and since there is no need for observer interaction, presentation of an identical data set to the method described above will always yield an identical result.

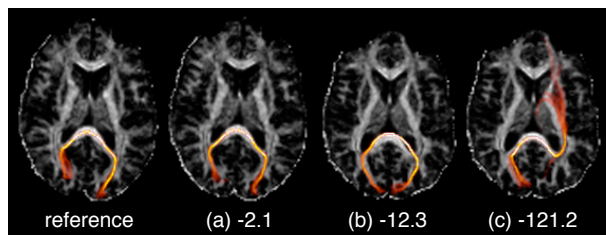
The present model-based approach to assessing tract similarity—which we first described in Clayden *et al.* (2007a), from which Figs 8.3–8.7 are taken—also has advantages over the heuristic method described in chapter 6. The first benefit is a general matter of principle: explicitly describing a tract matching model and its assumptions makes the method more scrutable than otherwise. Secondly, and more substantially, the median line representation of a tract can undergo affine transformation without complications; whereas the previously used field representation of a tract cannot be transformed without creating interpolation issues. This is helpful because it allows us to easily correct for gross head size or rotation differences between the reference and candidate tracts using standard affine image registration—as we have done above. Thirdly, the results from our previous approach to tract matching were quite strongly affected by the particular nature of the reference tract, and had a very narrow dynamic range. By contrast, Fig. 8.6 demonstrates that two very similar reference tracts do produce comparable—although not identical—results under the current model, while the matching probabilities assigned to dissimilar candidate tracts vary by orders of magnitude. Tracts (a), (b), (d) and (e) all represent appropriate matches to either splenium reference tract, and the fact that the best match under the original reference tract was also the best match under the alternative reference, out of a set of more than 200 candidates, does suggest a beneficial lack of sensitivity to small alterations in the reference tract.

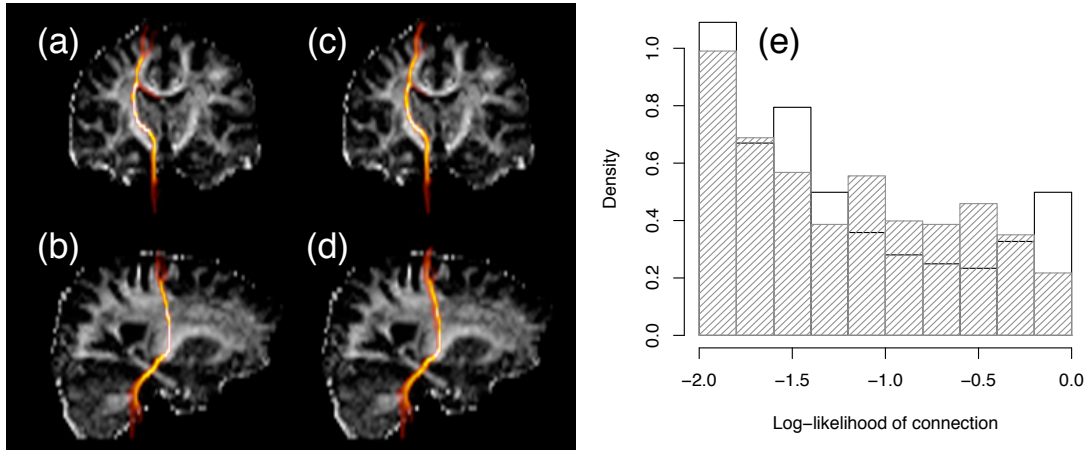
The greater dynamic range and probabilistic interpretation of the present approach to tract matching also suggest alternative uses for the likelihood data. Note that Eq. (8.8) describes a discrete matching distribution over a neighbourhood in each subject’s native space. The neighbourhood tractography method that we have employed so far is a maximum likelihood one, since we retain exactly the one tract which matches best under the model. For probabilistic tracts, the voxelwise likelihood of connection is then taken straight from the result of seeding at this single point. An alternative strategy is to find the expected value with respect to the matching distribution,

$$\hat{\phi}(\mathbf{x}) = \sum_{i=1}^N P(\mu = i | D) \phi_i(\mathbf{x}), \quad (8.11)$$

for each voxel location,  $\mathbf{x}$ , in the brain; thereby forming a weighted average field of connection likelihoods for any particular scan.

**Figure 8.7:** Log-ratios between matching likelihoods of the tracts shown and the matching likelihood of the reference tract. The reference tract has a log-ratio of zero by definition; (a) is the alternative reference tract; (b) is the best match in the novel candidate set; and (c) is the tract generated from the neighbourhood centre point.





**Figure 8.8:** Maximum likelihood corticospinal tract images in coronal (a) and sagittal (b) projections, and equivalent images for the weighted average tract data (c,d), all thresholded at the 1% level. Histograms of connection log-likelihoods for the maximum likelihood (plain) and weighted average (shaded) images are also shown (e).

Fig. 8.8 shows the effect of applying this strategy for the corticospinal tract example we looked at earlier. To save computation time, only the tracts with matching probabilities of greater than 0.01, of which there are five, were included in this weighted average. The images of the ML tract (a,b) and those of the weighted average tract (c,d) appear only subtly different, and their general trajectories are clearly very similar indeed. Nevertheless, there are notable differences. Firstly, the weighted average tract is wider than the ML one, giving more complete coverage of the voxels that are likely to represent the physical corticospinal tract. After thresholding at 1% there are 414 nonzero voxels in the average tract, as opposed to 321 in the ML version. As a result more data will be included in downstream tract-averaged comparisons of anisotropy across subjects, lending greater power to any statistical tests. Secondly, the distribution of connection likelihoods is markedly altered in the averaged tract, as shown by Fig. 8.8(e). It can be seen that the general trend in both tracts is for the larger connection likelihoods to occur less frequently, but in the ML tract—shown with unshaded bars—there is a significant upturn at the very top end of the range, representing an uncharacteristically large number of voxels that are very likely to be connected to the seed point. These are a direct result of the tight spatial distribution of streamlines near to the seed point, and are heavily seed point dependent (cf. Fig. 8.3). The average tract, on the other hand, incorporates data from several seed points and is therefore less affected by this problem. It can be seen from the histograms that the downward trend continues over the whole range of voxel values in this case. Only 7% of suprathreshold connection likelihood values are greater than 0.5 in the average tract, against 13% in the ML tract. It should be noted that calculating this kind of weighted average would be highly problematic using our earlier, heuristic similarity measure, owing to the very small differences in similarity that we found across sets of candidate tracts. Data from even very poorly matching tracts would consequently be well represented in the average tract.

Our model does have limitations, however. The median line cannot represent branches in the original set of streamlines; and as a result, the model cannot discriminate against such tracts, which may be considered desirable. (This, of course, will not be an issue in cases where the tractography algorithm produces a single streamline representation of a tract.) Secondly, the nature of Eq. (8.8) is such that the reference tract itself does not have the highest possible matching likelihood, and so the log-ratio calculated in Fig. 8.7 could be positive for some tracts. Moreover, since there is very little training data available for the length distributions, and so they are heavily affected by their regularisation terms, they do not fully compensate for the likelihood-increasing effects of the continuity cosines in very long tracts. Additionally, of course, any limitations and sensitivities to data quality that the chosen tractography algorithm

may have will apply in turn to our method.

The use of rois to constrain the paths that probabilistic streamlines may take (Conturo *et al.*, 1999; Heiervang *et al.*, 2006) is not precluded by our method. Indeed, a two-rois constraint could be applied to ameliorate problems with branches in the tracts if they proved significant, although we would advocate the avoidance of rois constraints wherever possible.

## 8.6 An unsupervised approach

We have shown that it is possible to capture the variability in shape and length between comparable tracts in different brain scans using a well-defined probabilistic model. However, the supervised approach that we have used up to now, whereby the model parameters are fitted using a group of training tracts chosen by hand, represents a rather suboptimal use of available information. We generated a small number of specialist training tracts, whilst some 200 candidate tracts were created for each test scan and then largely discarded. The hand-selection of training tracts also reintroduces an element of observer subjectivity into the process, albeit a reasonably minor one. On the other hand, if we could use the candidate tracts themselves to train the model whilst simultaneously finding a good match, then separate training data may not be required at all.

An unsupervised approach to the problem that uses the candidate tracts in this way could be constructed using EM, once again, with two generative models—one for matching tracts, and one for nonmatching tracts. We can then introduce a latent variable,  $z^i$ , indicating whether tract  $i$  matches the reference tract ( $z^i = 1$ ) or not ( $z^i = 0$ ). The “one best match” assumption that we have made up to this point can then be described by the equation

$$\sum_i P(z^i = 1) = 1. \quad (8.12)$$

Only one tract would therefore be drawn from the matching distribution, while all others are drawn from the nonmatching distribution. However, we introduce the additional possibility  $z^0 = 1$ , to mean that none of the candidate tracts represents a suitable match. Given an estimate for the model parameters,  $\hat{\omega}$ , the E-step of the algorithm would then involve calculating the posteriors

$$P(z^i = 1 | D) = \frac{P(z^i = 1) P(\mathbf{d}^i | \hat{\omega}, z^i = 1) \prod_{j \neq i} P(\mathbf{d}^j | z^j = 0)}{P(D)} \quad (8.13)$$

for each candidate tract—the likelihood of the tract in question under the matching model, multiplied by the likelihoods of all other tracts under the nonmatching model. The probability of no match among the candidates is given by

$$P(z^0 = 1 | D) = \frac{P(z^0 = 1) \prod_j P(\mathbf{d}^j | z^j = 0)}{P(D)}, \quad (8.14)$$

the normalised likelihood for every tract under the assumption that it does not match the reference. The evidence is

$$P(D) = \sum_i P(z^i = 1) P(\mathbf{d}^i | z^i = 1) \prod_{j \neq i} P(\mathbf{d}^j | z^j = 0) + P(z^0 = 1) \prod_j P(\mathbf{d}^j | z^j = 0). \quad (8.15)$$

The choice of priors for these calculations is not entirely straightforward. We may assume that each candidate tract is *a priori* equiprobable, say  $P(z^i = 1) = \gamma$  for  $i \in \{1..N\}$ , which then gives us  $P(z^0 = 1) = 1 - N\gamma$ . The difficulty is in the choice of  $\gamma$ , since it is hard to estimate in advance the chance of there being no match in the data. One option is to use  $\gamma = 1/(N + 1)$ , which makes the prior probability of no match the same as the prior for each candidate being the match, although this is an unprincipled position.

Our generative models for the matching and nonmatching tract data can be defined similarly to the single model that we used earlier. Since we found that the uniform component of

the distributions over similarity cosines tended to always shrink to zero, we can use—for the matching model—just a simple beta distribution for modelling similarity cosines. That is,

$$P(s_u^i | \alpha_u, z^i = 1) = \frac{\alpha_u}{2} \left( \frac{s_u^i + 1}{2} \right)^{\alpha_u - 1} \quad \forall i, u > 0.$$

The equivalent distributions for the nonmatching model can simply be uniform, and therefore quite independent of the reference tract. That is,

$$P(s_u^i | z^i = 0) = \frac{1}{2} \quad \forall i, u.$$

The length distributions remain multinomial for both models. The data likelihood under each model can therefore be written out as

$$P(\mathbf{d}^i | \boldsymbol{\omega}, z^i = 1) = P(L_1^i | L_1^*, z^i = 1) P(L_2^i | L_2^*, z^i = 1) \prod_{u=1}^{\check{L}_1^i} P(s_{-u}^i | \alpha_u, z^i = 1) \prod_{u=1}^{\check{L}_2^i} P(s_u^i | \alpha_u, z^i = 1), \quad (8.16)$$

and

$$P(\mathbf{d}^i | z^i = 0) = P(L_1^i | z^i = 0) P(L_2^i | z^i = 0) \left( \frac{1}{2} \right)^{\check{L}_1^i + \check{L}_2^i}, \quad (8.17)$$

where the parameter vector,  $\boldsymbol{\omega}$ , incorporates all of the  $\alpha_u$ .

We now step back from this mathematical deluge to discuss the meaning of these models in intuitive terms. As before, the matching tract is guided by the reference tract such that small deviations from the reference in its local direction are considered the most likely. The remaining tracts, which are not generated using the reference tract, use an uninformative distribution over similarity cosines, and so they may step in any direction with equal probability. Since we implicitly assumed that all unmatched tracts were equiprobable in our supervised method, this model is approximately analogous—although it does use (informative) multinomial distributions for the lengths. Beyond the end of the reference tract the two models effectively treat the tract in the same way, and so any contribution to the likelihoods from these regions will simply cancel out. They are therefore ignored in practice.

It follows from Eq. (8.13) that some tract,  $i$ , will be assigned a higher posterior matching probability than tract  $j$ , assuming equal priors, exactly when

$$\frac{P(\mathbf{d}^i | \boldsymbol{\omega}, z^i = 1)}{P(\mathbf{d}^i | z^i = 0)} > \frac{P(\mathbf{d}^j | \boldsymbol{\omega}, z^j = 1)}{P(\mathbf{d}^j | z^j = 0)}, \quad (8.18)$$

since the contributions from all other tracts cancel out. A suitable tract should therefore be a likely match, but also a relatively *unlikely* nonmatch. This makes sense since we are performing a model comparison; although, because only tract length affects the nonmatching likelihoods in the formulation we have given here, the impact of the nonmatching distributions will be small.

Eq. (8.18) does, however, explain why some other possible models are problematic. For example, it might seem more appropriate to use the continuity cosines to form a nonmatching model, so that the candidate tract is guided by itself in the absence of a reference tract. The problem, however, is that it is quite possible for a tract to be a good match to the reference *and* to be highly smooth; whereas, by Eq. (8.18), a smooth matching tract would be penalised relative to an unsmooth alternative using this form of nonmatching model. The implicit assumption of mutual exclusivity between the models is therefore not fulfilled. Hence, the continuity cosines are ignored altogether for present purposes.

The  $m$ -step of the algorithm is now relatively straightforward. The multinomial distributions can be updated as usual, using the matching posteriors as weights for each contributory tract length. The maximum likelihood estimator for  $\alpha$  in each similarity cosine distribution is given by

$$\hat{\alpha}_u = \frac{-2 \sum_{i>0} P(z^i = 1 | D)}{\sum_{i>0} P(z^i = 1 | D) \ln x_u^i}, \quad (8.19)$$

where

$$x_u^i = \left( \frac{s_{-u}^i + 1}{2} \right) \left( \frac{s_u^i + 1}{2} \right).$$

(In fact, Eq. (8.19) is not always quite accurate, since not all tracts will contribute a similarity cosine from both their left and right sides for every value of  $u$ ; but it conveys the intention.) In addition, since we wish to incorporate similarity cosine information from across a full data set, the sums over  $i$  in Eq. (8.19) will in practice be over all tracts for all subjects; although the  $\mathbf{E}$ -step above is performed for each volume individually.

At this stage we have a complete EM algorithm for unsupervised tract matching. There is, however, one outstanding issue. A consequence of the single best match assumption, Eq. (8.12), is that the final parameterisation of the model at convergence risks being very strongly customised to capture the characteristics of a small number of tracts, while matching all other tracts extremely poorly. We would expect this effect to be particularly noticeable when the number of contributing scans is comparable to the number of parameters in the model, since the algorithm then has a wide “choice” of tracts from which to select a small number of matches. To get around this issue, we can introduce a prior distribution over each  $\alpha$  parameter to regularise the likelihood function, and then take the maximum *a posteriori* estimate

$$\hat{\alpha} = \arg \max_{\alpha} \{ \ln P(\alpha | \mathbf{x}) \} = \arg \max_{\alpha} \left\{ \sum_i \ln P(x^i | \alpha) + \ln P(\alpha) \right\}.$$

For the prior, we use an exponential distribution with mean  $1/\lambda$ , defined by  $P(\alpha) = \lambda e^{-\lambda\alpha}$ . This prior will favour smaller values of alpha, thereby counteracting the upward tendency of model overfitting when there is little data available. The MAP estimator is then given by

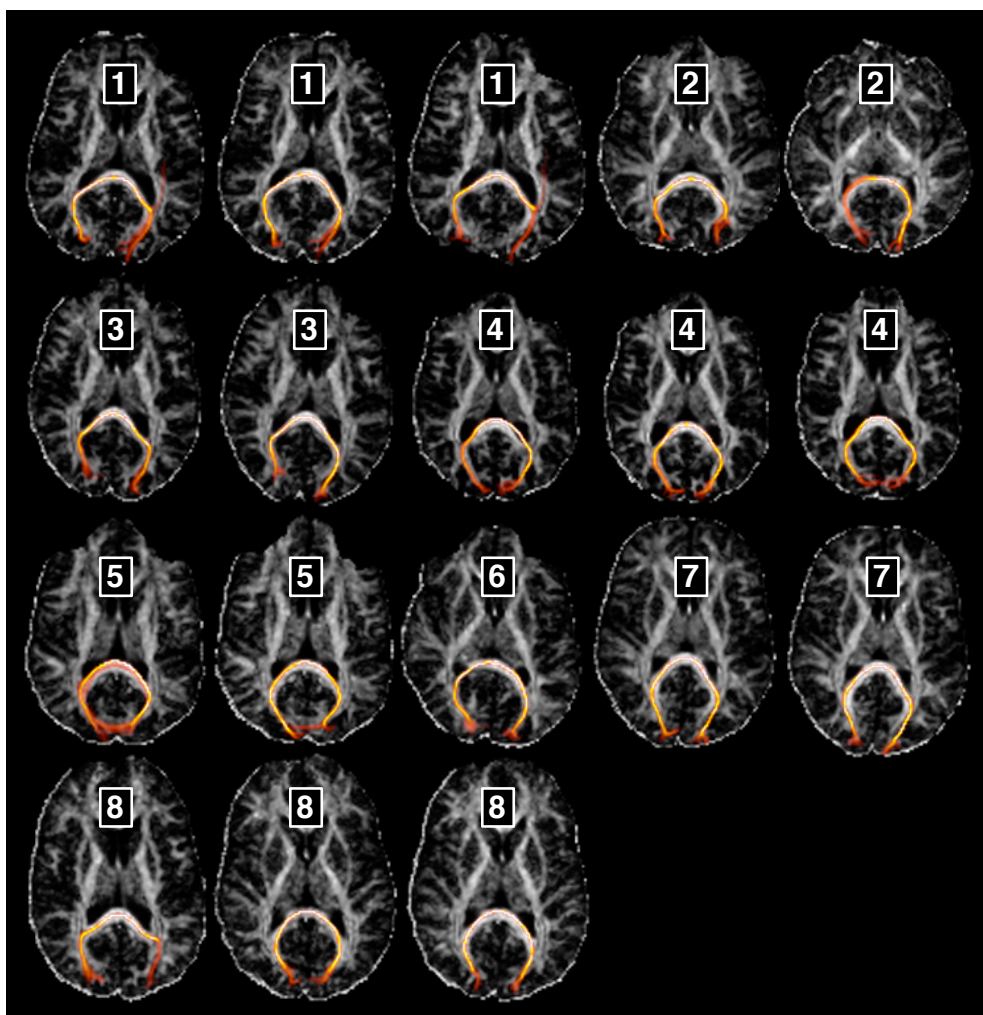
$$\hat{\alpha}_u = \frac{-2 \sum_{i>0} P(z^i = 1 | D)}{\sum_{i>0} P(z^i = 1 | D) \ln x_u^i - \lambda}. \quad (8.20)$$

Unlike Eq. (8.19), which is unbounded, Eq. (8.20) has an upper bound in the case where all the similarity cosines are maximal. Using a total of  $V$  volumes, and assuming that a good match can be found in each case—so that the null-match posterior,  $P(z^0 = 1)$ , is negligible—the numerator of Eq. (8.20) is approximately  $-2V$ , and so the upper bound will be given by  $2V/\lambda$ . Hence, the larger the number of brain volumes used for matching, the higher the bound and the smaller the impact that the prior distribution will have. This is appropriate since the risk of overfitting would also be diminished. We will take  $\lambda = 1$ .

We applied the technique to modelling and matching the corpus callosum splenium in 18 brain volumes collected from eight healthy young volunteers. The best matching tract in each volume under the resulting model is shown in Fig. 8.9. It can be seen that all tracts are plausible segmentations of the splenium; and there is also a high degree of topological similarity between tracts segmented from multiple scans of the same individual. Segmentations for subjects 1, 3, 4 and 7 are particularly alike between scans. This consistency is highly valuable for groupwise comparative analysis work. Posterior matching probabilities for these tracts ranged from 0.44 to greater than 0.99, using  $\lambda = 1$ . Without this regularisation, however, the posteriors were far greater, and in no case smaller than 0.98. In practice, these unregularised results are likely to be overly confident, due to the relatively small size of the data set.

There are a number of advantages of this method over the supervised approach. Firstly, of course, the removal of the need for training tracts allows a data set of any given size to be used to its fullest advantage, and reduces the time investment and subjectivity involved in creating a model for a particular tract. Only a reference tract need be defined *a priori*. Secondly, the existence of an explicit posterior probability of no match in a given volume is valuable. It should be stressed that this probability is conditional on assumptions implicit in the method and therefore care should be taken not to attach too much significance to it—but imposing thresholds on its value may nevertheless be a useful way to discard poor matches, or as an indication that the neighbourhood size should be increased. Indeed, the possibility exists of increasing the neighbourhood width incrementally until the null-match posterior drops below





**Figure 8.9:** Best matching splenium tracts from a full data set of 18 scans, selected using the EM approach to neighbourhood tractography and thresholded at 1%. The numbers indicate the subject number from which each scan was taken. The reference tract was taken from another scan of subject 3.

a certain level; and this can be done subject by subject since there is no requirement that each brain volume contribute equal numbers of candidate tracts.

The creation of a new model for each data set will be advantageous when one is dealing with, say, ageing brains. As we saw in chapter 7, there is evidence for greater variability among such brains, and this would be automatically allowed for by a model generated from an aged cohort. There is still the option of creating standardised models where this is deemed appropriate. Making the model more complex—most obviously by relaxing the assumption that the similarity cosine distributions are symmetric about the seed point, embodied by Eq. (8.9)—would also be possible for a large enough data set, and through its greater flexibility, this approach may result in even better matches.

The EM algorithm is not computationally demanding. It takes only around a minute to run using the 18 brain volumes from our experiment, and is expected to scale up linearly for larger data sets. Creating the set of B-spline tracts for each subject remains the most time-consuming part of the process, although this may be improved by reducing the number of probabilistic streamlines from which the median line is produced. Further testing would be needed to examine the impact of this kind of policy.

## 8.7 Conclusions

We have demonstrated in this chapter a formalised approach to neighbourhood tractography, whereby we explicitly represent the variability between subjects—relative to a reference tract—using probabilistic models, then learn parameters for those models, and finally use them for tract matching. We began with a supervised approach to model fitting, and then described a more complex variation that uses Expectation–Maximisation to learn appropriate parameters without the requirement for separate training data. Significantly, these models are able to allow for variability in the shape of candidate tracts in regions where it is most expected: particularly near where they terminate in grey matter.

We have not yet had the time to test these new NT techniques on clinical data sets, and this remains as future work. The results illustrated by Fig. 8.9 do, however, suggest that performance is considerably better than we obtained using the heuristic similarity measure (cf. Fig. 6.7). In the following chapter, we turn to look at a way to compare anisotropy—or other measures—downstream from the fibre tracking process, which does not simply involve averaging within the segmented region.



## Anisotropy profiling

IF TECHNIQUES such as neighbourhood tractography were to enable the robust segmentation of tracts representing equivalent fasciculi from a group of brain volumes, the question then arises, “What links or differentiates these tracts?” From a clinical perspective, we might be interested in looking for general differences in tract integrity between a healthy population and one affected by pathology. The work described in this chapter, which was completed under the supervision of Prof. David Laidlaw, attempts to look at integrity—as indicated by fractional anisotropy—on a fine-grained level, profiled along the length of a tract. The aim is to facilitate the testing of hypotheses about integrity at the within-tract level, and to investigate the behaviour and variability of anisotropy along a tract. This problem is separate to the one that neighbourhood tractography tries to solve, and is treated as such. We find evidence to suggest that although within-subject and within-group variance is large when FA is examined point-by-point, there can be sufficient regional differences between groups to ensure that subtle effects may well be masked by considering only mean FA values.

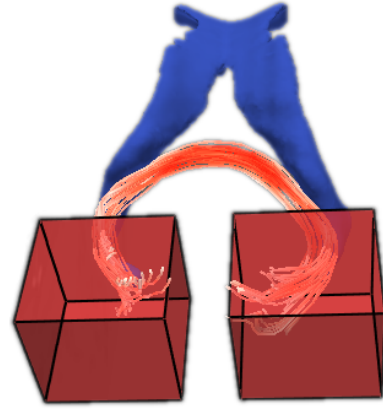
### 9.1 A single profile

To the extent that tractography is used at present for comparative clinical study, the most common approach is to average FA within the region segmented by the algorithm (e.g. Kanaan *et al.*, 2006), which may be represented by a line or a field. Region-averaged FA—however the region of interest is established—is a simple and useful way to study the effect of pathology on white matter integrity whilst controlling noise issues. On the other hand, ever greater numbers of studies are finding reduced FA effects in all kinds of pathologies, making such observations increasingly nonspecific; and since dMRI is the only available technique for studying structural white matter connectivity *in vivo*, independent corroboration or refutation of these results is extremely difficult. A partial list of scenarios in which reductions in FA have been observed could include schizophrenia (Ardekani *et al.*, 2003), multiple sclerosis (Ciccarelli *et al.*, 2003b), ischaemic leukoaraiosis in lacunar stroke (Jones *et al.*, 1999), epileptic patients after corpus callosotomy (Concha *et al.*, 2006), ischaemic stroke (Muñoz Maniega *et al.*, 2004) and normal ageing (O’Sullivan *et al.*, 2001).

For this study, six normal volunteers and five patients with vascular cognitive impairment (vci; a type of cognitive deficit which affects white matter and is quite common in elderly people) underwent a DTI protocol on a Siemens 1.5 T clinical scanner, with 12 noncollinear diffusion weighting gradient directions at a  $b$ -value of  $1000 \text{ s mm}^{-2}$ . The tractography infrastructure used for this work, BrainApp, uses a diffusion tensor-based deterministic streamlining algorithm, and visualises the results in terms of streamtubes and streamsurfaces (Zhang *et al.*, 2003). It uses whole brain seeding—which is possible in a reasonable time using a deterministic tractography algorithm—and thus avoids the selection constraint implicit to a neighbourhood or ROI seeding strategy. Simple streamline-based tractography lends itself very naturally to linear anisotropy profiling.

A streamtube is simply a piecewise linear streamline represented by a series of cylinders,

**Figure 9.1:** Example of a splenium streamtube set, segmented by placing a large region of interest near each end of the structure and retaining tubes passing through both. The shade of each streamtube indicates the local FA value. The blue structure represents the ventricles.

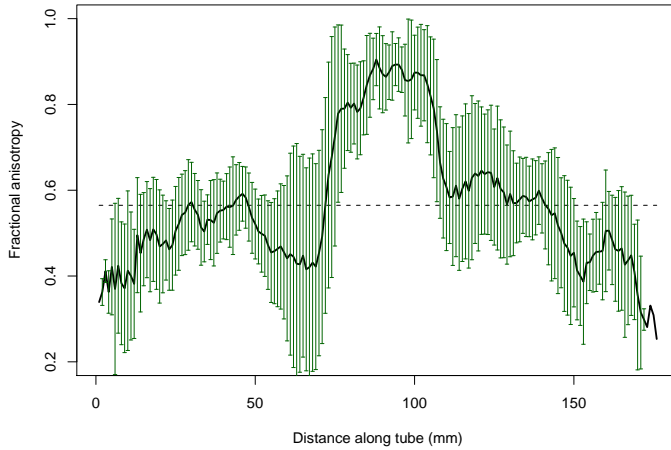


whose local radii may be constant or may be used to represent some characteristic of interest. A similar visualisation method has been used in other studies, such as Jones *et al.* (2005b). Working with tracts represented by single lines—rather than fields—is helpful for this work because it removes the need to linearise each tract before an anisotropy profile can be created. Ignoring its width, a streamtube,  $t_i$ , is therefore made up of piecewise linear line segments connecting a sequence of points,  $(\mathbf{p}_{i,a})$ , with  $a \in \{1..N_i\}$ , in the native acquisition space of the subject. The distance between successive points,  $d_i$ , is fixed in this space. Each of these tubes has a seed point, but unlike in the probability field output generated by `FSL ProbTrack`, the location of the seed point is not significant for the interpretation of the results, so we will not give it special treatment.

We first need to establish which tubes are of interest. Since BrainApp seeds throughout the brain, some kind of restriction is needed in order to focus on a specific white matter structure. Whatever method is chosen should be reproducible, however, so that it can be carried forward to comparative profiling between subjects. We used a two region of interest constraint to select the splenium of the corpus callosum, our tract of interest, with one ROI placed near the left end of the splenium tract and the other placed near the right end. These ROIs are symmetric, as per Conturo *et al.* (1999)—that is, they are treated identically, so swapping them would have no effect on the segmentation. This is not generally the case when one ROI provides the set of seed points, as in Abe *et al.* (2004) and some other studies. When working with streamtubes, this strategy amounts to taking the intersection of the set of tubes passing through ROI one, with the set of tubes passing through ROI two. An example of the result is shown in Fig. 9.1.

Since we have been critical of multiple ROI methods in earlier chapters, we will take a moment to justify this strategy. The important factors here are that the tractography algorithm being used to generate the streamtubes is deterministic, and that the seed points that generate the relevant streamtubes cannot be expected to form a compact neighbourhood, due to the whole brain seeding policy. Our objection about the effect of ROI constraints on interpretability due to the addition of extra conditional dependencies (cf. §6.5) only applies to output with a probabilistic significance. Constraining the algorithm by the selection of seed points is less relevant here; and neighbourhood tractography, which works on that principle, is not directly applicable. Ultimately, since the splenium is a coherent bundle with a distinctive shape, and is reasonably distinct from the rest of the corpus callosum and other nearby tracts in terms of the regions it connects together, the two ROI method is quite specific and reproducible enough. Moreover, it simply selects a set of streamtubes, just as choosing a number of seed points or clustering the streamtubes would. The effects are equivalent in essence.

Having “selected” the structure of interest, we can then plot the FA value,  $f_{i,a}$ , at each point on a tube,  $\mathbf{p}_{i,a}$ , thus forming an FA profile along the tube. (These values are interpolated from the FA data available at each voxel location.) Since all of these tubes are defined in the same space, aligning them is quite straightforward: we simply choose a plane which each streamtube must cross and consider the crossing points in each tube to be equivalent. We then examine the variability across the set of tubes at each point. This process produces a streamtube-averaged profile like the one shown in Fig. 9.2. In this case the distance between successive points,  $d_i$ ,



**Figure 9.2:** Pointwise mean FA along the set of splenium tubes segmented using the two ROI method in a single subject. The error bars indicate the mean plus or minus one standard deviation. The dashed line indicates the mean FA of the profile.

is 1 mm for all tubes. It should be noted that the alignment will handle differences in length well, but large shape differences, including kinks in some tubes, will render it inappropriate in some regions; and this is increasingly likely to occur as one moves away from the landmark plane.

Fig. 9.2 highlights two things in particular. Firstly, it is clear that the standard deviations are large to very large, relative to the means. Note, however, that on the left side of the graph in particular, the standard deviations are very large in a region near the middle of the tract, where the alignment plane was placed, and then shrink again further from the middle. This suggests that the variability is not primarily due to misalignment. One likely alternative cause is variation in the extent of partial volume effects. Some tubes will be nearer to the edge of the bundle than others, and the anisotropy at these locations is therefore more likely to be affected by proximity to grey matter or csf.

## 9.2 The median tube

Comparative profiling introduces some further issues. The questions of tube selection and alignment need to be reexamined, and differences in brain size must be compensated for in some way. We cannot simply use every tube selected in each brain, since the number of tubes selected is not fixed so bias would occur. We can't align tubes naïvely to a plane because each brain is represented in its own independent native space. And brain size cannot be neglected because it will affect the curvature of the structure and so the point homology.

Our approach to the first problem is to work only with the median tube from each brain; that is, the tube that minimises the average distance to all other tubes in the set. So, for a single subject whose splenium tube set contains  $N$  tubes, the median tube,  $t_m$ , is identified by calculating

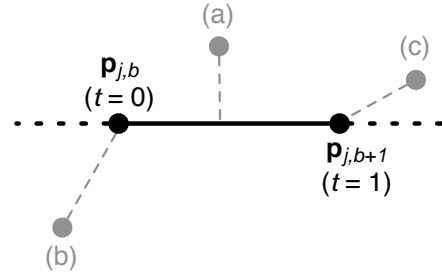
$$m = \arg \min_j \left\{ \frac{1}{N-1} \sum_{i=1; i \neq j}^N D(t_i, t_j) \right\}, \quad (9.1)$$

where  $D(t_i, t_j)$  is the distance between streamtube  $i$  and streamtube  $j$ , given by the average distance from the points on the longer tube to the shorter tube, viz.

$$D(t_i, t_j) = \begin{cases} \frac{1}{N_i} \sum_{a=1}^{N_i} d(\mathbf{p}_{i,a}, t_j) & \text{if } N_i \geq N_j \\ \frac{1}{N_j} \sum_{b=1}^{N_j} d(\mathbf{p}_{j,b}, t_i) & \text{otherwise.} \end{cases}$$

The point-to-tube distance,  $d(\mathbf{p}_{i,a}, t_j)$ , is given by the minimum distance between the point and a line segment delimited by successive points in tube  $t_j$ . The point-to-segment distance, in

**Figure 9.3:** The calculation of the distance,  $\hat{d}$ , between a point and line segment depends on whether the projection of the point onto the segment direction crosses the segment itself. In (a), the projection crosses the segment ( $0 < u < 1$ ), while in (b) and (c) it does not; and in these latter cases the shortest distances to the segment (dashed lines) are to one of its end points. The dotted extension of the line segment is shown for illustration.



turn, depends on the spatial arrangement of the point and segment (see Fig. 9.3). Mathematically, we parameterise the  $b$ th line segment as  $\mathbf{s}_{j,b}(t) = \mathbf{p}_{j,b} + t\mathbf{l}_{j,b}$ , where  $\mathbf{l}_{j,b} = \mathbf{p}_{j,b+1} - \mathbf{p}_{j,b}$  and  $t \in [0, 1]$ . The projection of the point  $\mathbf{p}_{i,a}$  onto the line segment—which forms the closest point between the two—is then given by  $\mathbf{p}_{j,b} + u\mathbf{l}_{j,b}$ , where

$$u = \frac{\mathbf{l}_{j,b} \cdot (\mathbf{p}_{i,a} - \mathbf{p}_{j,b})}{\mathbf{l}_{j,b} \cdot \mathbf{l}_{j,b}}. \quad (9.2)$$

The point and the line segment must, of course, be in the same space as one another. Now,  $u \in \mathbb{R}$ , and the distance between the point and the segment is calculated differently depending on whether the projection actually falls within the segment—i.e.  $u \in (0, 1)$ —or not. Specifically,

$$\hat{d}(\mathbf{p}_{i,a}, \mathbf{s}_{j,b}) = \begin{cases} \|\mathbf{p}_{i,a} - \mathbf{p}_{j,b}\| & \text{if } u \leq 0 \\ \|\mathbf{p}_{i,a} - (\mathbf{p}_{j,b} + u\mathbf{l}_{j,b})\| & \text{if } 0 < u < 1 \\ \|\mathbf{p}_{i,a} - \mathbf{p}_{j,b+1}\| & \text{if } u \geq 1. \end{cases} \quad (9.3)$$

We then have

$$d(\mathbf{p}_{i,a}, t_j) = \min_b \{\hat{d}(\mathbf{p}_{i,a}, \mathbf{s}_{j,b})\}. \quad (9.4)$$

This is a standard formulation of the distance between a point and a line segment, but unfortunately it is a case in which the maths makes a simple concept look complicated. Eq. (9.2) is mathematical infrastructure for Eq. (9.3), which embodies the fact that if the line orthogonal to the line segment and passing through the point  $\mathbf{p}_{i,a}$  does not cross the line segment, then the nearest point on the segment is in fact one of the end points. Fig. 9.3 illustrates this, for all three cases in Eq. (9.3). Note that if the next line segment, from  $\mathbf{p}_{j,b+1}$  to  $\mathbf{p}_{j,b+2}$ , were to be collinear with the one illustrated, then the distance from point (c) to that segment would be lower than the distance shown, affecting the value of Eq. (9.4) appropriately.

Thus—finally—Eq. (9.1) is fully defined, and we can find the median tube in this way for each subject. This arrangement has the advantage that the median will tend to be towards the physical centre of a bundle of tubes, and therefore any partial volume effects should be relatively small. Incidentally, this justification differs slightly from that given for using the median streamline for tract matching in chapter 8, where the median was used simply because it epitomises the shape of a set of streamlines.

### 9.3 Intersubject tube alignment

As we have already mentioned, the tube sets representing the splenium of each subject's corpus callosum are necessarily each defined in their own space; and so absolute point locations are not directly comparable between subjects. In order to work around this complication, we observe that the splenium, being an interhemispheric fasciculus, always crosses the brain's midsagittal divide. (In fact, the placement of the rois guarantees this, since one is in the left hemisphere and one in the right.) This divide can be acceptably approximated by a plane. A number of methods have been proposed for automatically extracting this plane (e.g. Hu & Nowinski, 2003; Volkau *et al.*, 2006), but for this work we established its location in each subject manually, by placing four points,  $\mathbf{r}_1$  to  $\mathbf{r}_4$ , on the midsagittal divide by eye—thus marking the corners of a trapezium. Since three points are sufficient to establish a plane, the distance of the

fourth point to the plane was used as a simple error measurement to gauge the consistency of the placement. This distance is given by

$$\delta = \mathbf{n} \cdot (\mathbf{r}_4 - \mathbf{r}_1) ,$$

where

$$\mathbf{n} = \frac{(\mathbf{r}_2 - \mathbf{r}_1) \wedge (\mathbf{r}_3 - \mathbf{r}_1)}{\|(\mathbf{r}_2 - \mathbf{r}_1) \wedge (\mathbf{r}_3 - \mathbf{r}_1)\|} ,$$

the unit vector normal to the plane on which the points  $\mathbf{r}_1$  to  $\mathbf{r}_3$  lie. The mean placement error across all subjects,  $\langle \delta \rangle$ , was 0.90 mm.

Having established this midplane, we can find the location where each subject's median tube crosses the plane by first working out in which line segment the crossing occurs, and then finding the exact intersection of that segment with the plane. If the relevant line segment passes through the points  $\mathbf{r}_5$  and  $\mathbf{r}_6$ , it can be expressed as

$$\mathbf{s}(t) = \mathbf{r}_5 + t(\mathbf{r}_6 - \mathbf{r}_5) ,$$

and a bit more geometry yields the value of  $t$  where the line segment crosses the plane to be given by the ratio of determinants

$$t = - \frac{\det \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_5 \\ y_1 & y_2 & y_3 & y_5 \\ z_1 & z_2 & z_3 & z_5 \end{bmatrix}}{\det \begin{bmatrix} 1 & 1 & 1 & 0 \\ x_1 & x_2 & x_3 & x_6 - x_5 \\ y_1 & y_2 & y_3 & y_6 - y_5 \\ z_1 & z_2 & z_3 & z_6 - z_5 \end{bmatrix}} ,$$

where  $\mathbf{r}_1 = (x_1, y_1, z_1)$  and so on. We then translate the co-ordinate system of each native space so that this intersection point is at the origin. Finally, on the assumption that the point where the median tube crosses the midsagittal divide is equivalent across brains, we treat all of these translated spaces as being equivalent. It is now possible to combine the median tubes from all subjects into an *intersubject* tube set, and find an intersubject median tube from this set.

Correcting for translational differences between subjects is not sufficient, however, since the shapes of the different subjects' spleniums will still vary due to differences in brain size. One approach to this problem is to use the intersubject median-of-medians tube,  $t_M$ , as a spatial reference, and take an FA value,  $f'$ , for each tube at each point on this median by finding the nearest neighbour point on each separate tube. That is,

$$f'_{i,a} = f_{i,\tilde{b}} ,$$

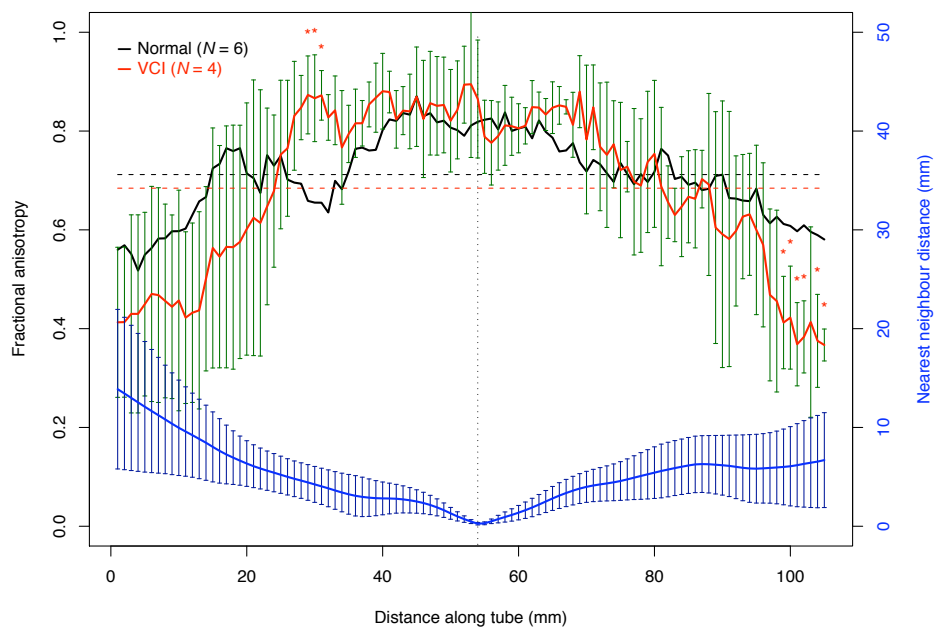
where

$$\tilde{b} = \arg \min_b \| \mathbf{p}_{M,a} - \mathbf{p}_{i,b} \| ;$$

so the  $a$ th FA value from tube  $i$  is the value at that point on  $t_i$  that is closest to the  $a$ th point on  $t_M$ , with  $i$  now indexing over subjects. This gives us a one-dimensional FA profile of fixed length for all subjects.

## 9.4 Comparative profiling

Fig. 9.4 shows the result of performing the whole process described above on a full data set. We located the splenium, using the two ROI method, in each subject. We then calculated a single intersubject median tube by combining all subjects' individual medians together; but subsequently separated them into patient and control groups once more for generating the averaged profiles shown in the figure. The intersubject median's FA data was not included,



**Figure 9.4:** An example of comparative profiling between groups of subjects. The red line with green error bars shows the average (plus or minus one standard deviation) value of  $f'$ , averaged across all subjects with VCI, at each point on the intersubject median tube. The black line shows the mean across the normal subjects. Appropriately coloured horizontal dashed lines show the profile mean FA. The blue line with blue error bars indicates the mean (plus or minus one standard deviation) distance from the intersubject median tube to its nearest neighbour at each point, across all subjects. The vertical dotted line shows the location of the midplane.

reducing the number of vCI subjects contributing to four. These initial results were first presented in Clayden *et al.* (2007b).

In Fig. 9.4, red stars indicate significant ( $P < 0.05$ ) differences between the groups using a two tailed  $t$ -test on  $f'$  data at each point. Since no correction for multiple comparisons was performed, these differences are tentative results at best, but they are somewhat informative nevertheless. Since the significant points are clustered into two (almost) contiguous regions, it seems unlikely that the differences are due to random noise effects; although the combination of interpolation and the nearest neighbour process makes successive points somewhat interdependent. It is interesting to note that while the grand mean  $FA$ , indicated by dashed horizontal lines, is lower for the vCI group than for the normal group—although this difference was not significant—the two regions differing between the profiles are not consistent in the sign of the difference between the groups. The region at the right hand end of Fig. 9.4 shows lower  $FA$  in the vCI group, which is the most common finding in pathological groups, while the region on the left side of the graph shows *higher*  $FA$  in vCI. This may be because the region represents an area of crossing fibres. If one of the two fibre populations were to preferentially suffer a loss of integrity, an increase in  $FA$  would be expected. To the left of the significant region,  $FA$  is decreased relative to the normal population again, although the error bars are too large for this to be significant.

The large nearest neighbour distances in this latter region may be responsible for the large variability which is particularly noticeable at the left hand end of the profiles. The blue curve indicates the mean and standard deviation of the distances from the intersubject median tube to each subject's individual median. This is zero by definition at the midplane—indicated with a vertical dotted line—and tends to increase as one gets further from there. The greater this distance becomes, the greater the divergence of the median tubes from one another; but it is not clear whether, or to what extent, an increase in divergence makes the profiled  $FA$  values intrinsically less comparable.

## 9.5 Discussion

The approach to anisotropy profiling described above has allowed us to explore some of the issues involved with this kind of comparative analysis, and to get a sense of the variability in anisotropy along a major tract. There are, however, evident reasons that this technique would not be very widely useful in its current state. Firstly, not all tracts in the brain cross the midsagittal divide, so using this landmark for intersubject alignment will not be possible in all cases. Secondly, the use of nearest neighbours for establishing a point homology between tubes is not robust, and the performance of the technique will depend on the shape of the tract of interest. One possible way of avoiding both of these issues is to use registration for alignment of median tubes between subjects. This would solve the problem of handling differences in brain size at the same time as annulling translational misalignment. It would be less tract-specific than the combination of techniques described in §9.3, and so if it worked well enough it would be applicable, in theory, to any tract of interest in the brain. Another possible avenue would be to use the probabilistic neighbourhood tractography methods described in chapter 8 to select a representative line for each subject, rather than taking the median. This would circumvent the limitations of the two ROI method in more complex tracts than the splenium.

In addition to dealing with these systematic limitations, we would need to apply the profiling process to more data to get a clearer picture of its effectiveness, or to draw any serious clinical conclusions. In particular, it would be interesting to study differences in the profiles between scans of a single subject, and between two normal populations. We would also need to look at other tracts. It may be that the full  $FA$  profile is actually too noisy a representation to be generally useful; but it is nevertheless suggested by the results so far that the mean  $FA$  along a streamtube, or group of tubes, is only a perfunctory summary of the available information. Fig. 9.2 shows that even allowing for large error bounds the  $FA$  along a tract in a single subject is not well encapsulated by the mean, and Fig. 9.4 demonstrates potential regions of difference between healthy and possibly abnormal profiles despite there being no significant difference in the means. We have also done some work in which the profiles were filtered for high

frequency noise by convolving them with a Gaussian smoothing kernel—which has not been shown here—but it remains unclear whether or not this would be beneficial. It may be that one could use this kind of smoothing to make multiple comparisons correction less of a problem, as *vbm* does, but the choice of variance for the smoothing kernel might be hard to justify. All of this is left as future work.

The ability to meaningfully compare anisotropy—or diffusivity, or any other measure of interest—between groups at a fine-grained but tract-specific level could be very useful for comparative analysis in white matter, but for the moment there are, as we have discussed, a number of hurdles in the way of the profiling approach we have described here.



# Conclusions

THE AIMS of this thesis were to develop methods to facilitate the robust segmentation of specific white matter structures from multiple dMRI brain volumes, and the subsequent comparative analysis of the segmented regions. In this final chapter we review the extent to which the work described in the previous chapters has met these aims, and discuss the work that still needs to be done.

## 10.1 Tract segmentation

As we described in the introduction, the study of structural human brain connectivity *in vivo* really only began with the invention of DTI in the mid-1990s. Over the course of less than a decade, since the possibility of using tensor-derived metrics to probe white matter integrity took hold, a sizeable clinical literature based on the method has amassed; but the techniques are still quite immature. Ideally, one would begin studying a disease in which a loss of connective efficacy is a suspected factor by applying a whole-brain analysis technique such as VBM to suggest regions of localised contrast between patient and control populations. A replication study might then hope to characterise the effect on any implicated white matter structures more clearly, and look for evidence that particular white matter degradation is specifically linked to the pathology in question. Unfortunately the reality is less straightforward.

When applied to maps of diffusion anisotropy, the VBM method is not robust. As we discussed in §6.1, the choice of smoothing kernel can have a very substantial effect on the results—not just quantitatively, but qualitatively too, with regions of contrast appearing in quite different brain areas as the kernel width is altered. Since this parameter of the method is usually chosen for each individual study without recourse to any firm principles, the scope for spurious and misleading results is unsettling. Moreover, it is easy—although unwise—to forget that FA itself has limitations as a proxy for integrity. We discussed in §9.4 that FA would be expected to *increase* if one of a pair of crossing fibre populations were to be preferentially degraded, and therefore it cannot necessarily be trusted as a reliable indicator of disease in crossing fibre regions. Even if this shortcoming did not exist, the expressiveness of a single scalar parameter will always be limited.

Tract-based spatial statistics may in practice take over as the method of choice when looking for localised differences between populations with limited or no prior knowledge. Its only overt parameter is an FA threshold that is applied to the skeletonised anisotropy maps, which will usually affect the results only quantitatively due to its impact on the number of voxels surviving to the multiple comparisons correction stage. The technique does perform time-consuming nonlinear registration of each subject’s brain volume to every other, which makes it scale badly to large data sets, but the introduction of a standardised template skeleton might be possible to remove this issue. TBSS is certainly an attractive approach, although it is not truly “tract-based” since the skeletonisation process will find any ridge in the anisotropy map and has no concept of white matter structure or connectivity. For most purposes, the approximation is however an adequate one.

On the other hand, automated methods for tract-specific segmentation and comparative analysis are more or less nonexistent. Regions of interest can be defined in standard space and then transferred to native space using registration, and used to constrain tractography; but this multiple ROI approach has a number of drawbacks, as we discussed in §6.5. Like any registration-based transformation, this one will engender some inaccuracy in the placement of the ROIs in native space; but in any case these ROIs encode prior knowledge about the topology of tracts in an unintuitive manner, which is informed primarily by experience with tractography rather than direct knowledge of anatomy. The use of “termination” and “removal” masks by Heiervang *et al.* (2006), for example, is presumably founded on past experience, during which the authors observed some pathways straying into these regions and deemed them aberrant or undesirable. The problem is that the ROIs might need to be redrawn for use with a different tractography algorithm.

We would argue that our representation of prior knowledge about tract topology in terms of reference tracts is more intuitive, more transferrable and ultimately more reliable. Information about the expected route of the tract is given along its entire length, but this richer prior information is not used to directly constrain the fibre tracking algorithm—rather, it guides the choice of tractography results from among a number of candidate seed points. The combined process of matching tracts to a reference and choosing a segmentation *a posteriori* based on these matches is neighbourhood tractography, a largely automated approach that we have invented and refined over the course of the thesis.

In chapter 6, we described a heuristic similarity measure for matching tracts and outlined the principle of NT. We demonstrated that the method improved segmentation consistency over a naïve alternative method in a group of healthy volunteers; and then, in chapter 7, we found similar benefits in a healthy aged cohort. We were able to use a reference tract taken from an aged brain to successfully select tracts from the younger group, and thereby to show anisotropy differences between the groups in a specific tract where previous whole-brain studies have suggested that one might be present. We have also discussed how standardised reference tracts can be generated from a white matter atlas, and used these references in a practical study.

To ameliorate some of the shortcomings of our simple first approach to tract matching for NT, we reformulated the problem in formal probabilistic terms in chapter 8, and took a machine learning perspective toward its solution. The models that we used to represent the relationships between matching tracts were parameterised such that the extent of deviation from the route of the reference tract can vary along its length, meaning that large variability within the data used to fit the model will result in only small penalties for straying from the reference. To learn suitable parameters, we initially took a supervised maximum likelihood approach, in which a group of training tracts is selected by hand in addition to the reference tract; but later showed that an EM algorithm could be used to successfully find matches in a data set without a separate training phase.

The main parameter of NT methods is the neighbourhood width. If this is set too small then no appropriate match to the reference will be found, and if it is set too large then the process will take a very long time to run. The limiting case of seeding throughout the brain is theoretically optimal in the sense that if a matching tract can be produced then it should be found this way—unless there happens to be another fasciculus with very similar shape and length in another part of the brain—but the practical consideration of run time makes this an unwise strategy. In any case, the use of tract similarity measures or matching models gives us an indication of the acceptability of the best match that we can use to our advantage. As we mentioned in §8.6, the null-match posterior probability that is available in the unsupervised probabilistic case could be used as the basis of a rejection criterion. To minimise run time, the neighbourhood width could be chosen separately for each brain volume, being increased in steps until the null-match posterior drops below a certain level. Nonetheless, a proper analysis of the effect of neighbourhood width would be a useful avenue for future work.

Other parameters arising in the model-based NT methods, such as the residual error threshold,  $\eta$ , and the streamline length quantile,  $\xi$ , may also have some effect on the outcome. But the former is relevant only to the reference tract, and we have found no reason to vary the latter from one brain volume to another, so in practice there should be little reason for them to

vary between studies and therefore become a point of weakness in any results.

We have not yet had time to apply the probabilistic model-based variants of NT to clinical data sets, or to develop atlas-based reference tracts for use with them; and these are important areas for future work. With them in place, however, we feel that the approach could represent a useful, robust and automated technique for the segmentation of specific tracts.

## 10.2 Comparative analysis

Once similar regions are segmented from a number of brain volumes, the simplest approach to comparative analysis between groups is to average a scalar measure of interest within each region and statistically compare the range of values thus obtained. This average can be weighted using the voxelwise likelihoods of connection to the seed point produced by an algorithm such as FSL ProbTrack—as we did in chapter 7. Using probabilistic neighbourhood tractography, we can also include data derived from multiple seed points, weighting according to the corresponding matching posteriors as in §8.5. It would be constructive to examine the benefits (or otherwise) of these weighting schemes more closely than we have done above.

In chapter 9, we explored the possibility of profiling anisotropy along tracts rather than simply averaging its value within the regions representing the relevant fasciculus. This raises some difficult questions about point homology in different brain volumes, but our initial results nevertheless suggest that this kind of approach may be able to yield some additional meaningful information.

Ultimately, comparing anisotropy between populations—however it is done—is only going to take *in vivo* white matter studies so far. Mean diffusivity is a mathematically independent measure for characterising diffusion, but in practice it is generally negatively correlated with FA. Combination of diffusion data with information from other magnetic resonance methods may prove more fruitful: spatially localised brain “activation” data from functional MRI, or metrics derived from magnetisation transfer imaging (which was briefly mentioned in §4.5) may help, if the concomitant coregistration issues can be worked out. Even more broadly, there is scope for incorporating data from fields such as genetics into advanced studies.

## 10.3 Final remarks

It is our hope that the methodological developments set out in this thesis will be helpful for ongoing work investigating whether connective changes are systematically linked to outwardly visible pathology. We believe that we have made useful progress towards robust segmentation of tracts of interest, which, so long as it remained problematic, has been a distracting prerequisite for meaningful investigation of the differences and similarities between comparable white matter structures.

As its resolution and noise properties improve, the potential of dMRI should continue to increase, although these developments will probably bring new challenges as well. Methods for examining connectivity may need to become more sophisticated, but a definitive test of the disconnection hypothesis could be at hand in the foreseeable future.



# A

## List of abbreviations

The following table describes the significance of various abbreviations and acronyms used throughout the thesis. The page on which each one is defined or first used is also given. Nonstandard acronyms that are used in some places to abbreviate the names of certain white matter structures are not in this list.

Abbreviation	Meaning	Page
AVF	anisotropic volume fraction	51
BEDPOST	Bayesian Estimation of Diffusion Parameters Obtained using Sampling Techniques	51
c.d.f.	cumulative density function	11
CSF	cerebrospinal fluid	4
CV	coefficient of variation	71
dmRI	diffusion magnetic resonance imaging	3
DSI	diffusion spectrum imaging	56
DTI	diffusion tensor imaging	40
EM	Expectation–Maximisation	16
EPI	echo planar imaging	34
FA	fractional anisotropy	41
FACT	Fibre Assignment by Continuous Tracking	47
FLIRT	FMRIB’s Linear Image Registration Tool	69
fmRI	functional magnetic resonance imaging	45
FMRIB	Centre for Functional Magnetic Resonance Imaging of the Brain (Oxford, UK)	53
FMT	fast marching tractography	53
FSL	FMRIB software library	53
FID	free induction decay	31
FWHM	full width at half maximum	62
HARDI	high angular resolution diffusion imaging	54
i.i.d.	independent and identically distributed	14
MAP	maximum <i>a posteriori</i>	16
MC	Monte Carlo	19
MCMC	Markov chain Monte Carlo	22
MD	mean diffusivity	41
ML	maximum likelihood	14
MNI	Montréal Neurological Institute	69
MRI	magnetic resonance imaging	3
NMR	nuclear magnetic resonance	4

*continued on next page*

<b>Abbreviation</b>	<b>Meaning</b>	<b>Page</b>
NT	neighbourhood tractography	64
ODF	orientation distribution function	56
PAS	persistent angular structure	56
p.d.f.	probability density function	11
p.m.f.	probability mass function	10
RF	radiofrequency	29
ROI	region of interest	63
TBSS	tract-based spatial statistics	62
TE	echo time	32
TI	inversion time	32
TR	repetition time	32
VBM	voxel-based morphometry	61
VCI	vascular cognitive impairment	103

# B

---

## Summary of publications

The following is a list of works published by the author during the course of the doctorate. Many of these works are cited in the text and therefore also appear in the full bibliography which follows.

### Journal papers

- M.E. Bastin, **J.D. Clayden**, A. Pattie, I.F. Gerrish, J.M. Wardlaw & I.J. Deary (in press). Diffusion tensor and magnetization transfer MRI measurements of periventricular white matter hyperintensities in old age. To appear in *Neurobiology of Aging*.
- **J.D. Clayden**, A.J. Storkey & M.E. Bastin (2007). A probabilistic model-based approach to consistent white matter tract segmentation. *IEEE Transactions on Medical Imaging* 26(11):1555–1561.
- **J.D. Clayden**, M.E. Bastin & A.J. Storkey (2006). Improved segmentation reproducibility in group tractography using a quantitative tract similarity measure. *NeuroImage* 33(2):482–492.
- I.J. Deary, M.E. Bastin, A. Pattie, **J.D. Clayden**, L.J. Whalley, J.M. Starr & J.M. Wardlaw (2006). White matter integrity and cognition in childhood and old age. *Neurology* 66(4):505–512.

### Peer-reviewed conference abstracts

- **J.D. Clayden**, A.J. Storkey & M.E. Bastin (2007). A probabilistic model-based approach to consistent white matter tract segmentation. In *Proceedings of the ISMRM-ESMRMB Joint Annual Meeting*, p. 78.
- **J.D. Clayden**, S. Zhang, S. Correia & D.H. Laidlaw (2007). Fine-grained comparison of anisotropy differences between groups of white matter tracts. In *Proceedings of the ISMRM-ESMRMB Joint Annual Meeting*, p. 82.
- M.E. Bastin, **J.D. Clayden**, A. Pattie, I.F. Gerrish, J.M. Wardlaw & I.J. Deary (2007). DTI and MTI parameters correlate in periventricular white matter hyperintensities in old age. In *Proceedings of the ISMRM-ESMRMB Joint Annual Meeting*, p. 2160.
- **J.D. Clayden**, M.E. Bastin & A.J. Storkey (2006). Neighbourhood tractography: a new approach to seed point placement for group fibre tracking. In *Proceedings of the British Chapter of the ISMRM*, p. P34.
- **J.D. Clayden**, M.E. Bastin & A.J. Storkey (2006). Automated assessment of tract similarity in group diffusion MRI data. In *Proceedings of the ISMRM 14th Scientific Meeting & Exhibition*, p. 2742.

- **J.D. Clayden**, D.K.S. Marjoram, M.E. Bastin, E.C. Johnstone & S.M. Lawrie (2005). Towards an automated method for white matter integrity comparison between populations. In *Proceedings of the ESMRMB 22nd Annual Meeting*, p. 508.
- I.J. Deary, M.E. Bastin, A. Pattie, **J.D. Clayden** & J.M. Wardlaw (2005). Cognitive ageing and white matter integrity: a diffusion tensor and magnetisation transfer MRI study. In *Proceedings of the ISMRM 13th Scientific Meeting & Exhibition*.
- I.J. Deary, M.E. Bastin, A. Pattie, **J.D. Clayden**, J.M. Starr & J.M. Wardlaw (2005). White matter integrity and cognition in childhood and old age: a diffusion tensor MRI study. *Methods for Quantitative MRI of Human Brain*. ISMRM Workshop, Lake Louise, Canada.



---

# Bibliography

- Abe O., Aoki S., Hayashi N., Yamada H., Kunimatsu A., Mori H., Yoshikawa T., Okubo T. & Ohtomo K. (2002). Normal aging in the central nervous system: quantitative MR diffusion-tensor analysis. *Neurobiology of Aging* **23**:433–441.
- Abe O., Yamada H., Masutani Y., Aoki S., Kunimatsu A., Yamasue H., Fukuda R., Kasai K., Hayashi N., Masumoto T., Mori H., Soma T. & Ohtomo K. (2004). Amyotrophic lateral sclerosis: diffusion tensor tractography and voxel-based analysis. *NMR in Biomedicine* **17**(6):411–416.
- Alexander A., Hasan K., Lazar M., Tsuruda J. & Parker D. (2001). Analysis of partial volume effects in diffusion-tensor MRI. *Magnetic Resonance in Medicine* **45**(5):770–780.
- Alexander D. (2005). Maximum entropy spherical deconvolution for diffusion MRI. In *Information Processing in Medical Imaging*, vol. 3565 of *Lecture Notes in Computer Science*, pp. 76–87. Springer-Verlag.
- Alexander D., Barker G. & Arridge S. (2002). Detection and modeling of non-Gaussian apparent diffusion coefficient profiles in human brain data. *Magnetic Resonance in Medicine* **48**(2):331–340.
- Andrieu C., Freitas N., Doucet A. & Jordan M. (2003). An introduction to MCMC for machine learning. *Machine Learning* **50**:5–43.
- Anjari M., Srinivasan L., Allsop J., Hajnal J., Rutherford M., Edwards A. & Counsell S. (2007). Diffusion tensor imaging with tract-based spatial statistics reveals local white matter abnormalities in preterm infants. *NeuroImage* **35**:1021–1027.
- Ardekani B., Nierenberg J., Hoptman M., Javitt D. & Lim K. (2003). MRI study of white matter diffusion anisotropy in schizophrenia. *Neuroreport* **14**:2025–2029.
- Ashburner J. & Friston K. (2000). Voxel-based morphometry—the methods. *NeuroImage* **11**(6):805–821.
- Ashburner J. & Friston K. (2001). Why voxel-based morphometry should be used. *NeuroImage* **14**:1238–1243.
- Assaf Y., Ben-Bashat D., Chapman J., Peled S., Biton I., Kafri M., Segev Y., Hendler T., Korczyn A., Graif M. & Cohen Y. (2002). High b-value q-space analyzed diffusion-weighted MRI: application to multiple sclerosis. *Magnetic Resonance in Medicine* **47**(1):115–126.
- Assaf Y., Chapman J., Ben-Bashat D., Hendler T., Segev Y., Korczyn A., Graif M. & Cohen Y. (2005). White matter changes in multiple sclerosis: correlation of q-space diffusion MRI and <sup>1</sup>H MRS. *Magnetic Resonance Imaging* **23**(6):703–710.

- Baird A. & Warach S. (1998). Magnetic resonance imaging of acute stroke. *Journal of Cerebral Blood Flow & Metabolism* **18**(6):583–609.
- Basser P., Mattiello J. & Le Bihan D. (1994a). Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of Magnetic Resonance, Series B* **103**(3):247–254.
- Basser P., Mattiello J. & Le Bihan D. (1994b). MR diffusion tensor spectroscopy and imaging. *Biophysical Journal* **66**(1):259–267.
- Basser P., Pajevic S., Pierpaoli C., Duda J. & Aldroubi A. (2000). In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine* **44**(4):625–632.
- Basser P. & Pierpaoli C. (1996). Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *Journal of Magnetic Resonance, Series B* **111**(3):209–219.
- Bastin M., Clayden J., Pattie A., Gerrish I., Wardlaw J. & Deary I. (2007). Diffusion tensor and magnetization transfer MRI measurements of periventricular white matter hyperintensities in old age. *Neurobiology of Aging* **in press**.
- Batchelor P., Atkinson D., Hill D., Calamante F. & Connelly A. (2003). Anisotropic noise propagation in diffusion tensor MRI sampling schemes. *Magnetic Resonance in Medicine* **49**(6):1143–1151.
- Batchelor P., Calamante F., Tournier J.D., Atkinson D., Hill D. & Connelly A. (2006). Quantification of the shape of fiber tracts. *Magnetic Resonance in Medicine* **55**:894–903.
- Behrens T., Johansen-Berg H., Jbabdi S., Rushworth M. & Woolrich M. (2007). Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage* **34**(1):144–155.
- Behrens T., Johansen-Berg H., Woolrich M., Smith S., Wheeler-Kingshott C., Boulby P., Barker G., Sillery E., Sheehan K., Ciccarelli O., Thompson A., Brady J. & Matthews P. (2003a). Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nature Neuroscience* **6**(7):750–757.
- Behrens T., Woolrich M., Jenkinson M., Johansen-Berg H., Nunes R., Clare S., Matthews P., Brady J. & Smith S. (2003b). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine* **50**(5):1077–1088.
- Bishop C. (2006). *Pattern recognition and machine learning*. Springer, New York.
- Biton I., Duncan I. & Cohen Y. (2006). High b-value q-space diffusion MRI in myelin-deficient rat spinal cords. *Magnetic Resonance Imaging* **24**:161–166.
- Bloch F. (1946). Nuclear induction. *Physics Review* **70**(7-8):460–474.
- Böhm W., Farin G. & Kahmann J. (1984). A survey of curve and surface methods in CAGD. *Computer Aided Geometric Design* **1**(1):1–60.
- Bransden B. & Joachain C. (1989). *Introduction to quantum mechanics*. Longman.
- Brun A., Knutsson H., Park H.J., Shenton M. & Westin C.F. (2004). Clustering fiber traces using normalized cuts. In *Medical Image Computing and Computer-Assisted Intervention*, vol. 3216 of *Lecture Notes in Computer Science*, pp. 368–375. Springer-Verlag.
- Bürgel U., Amunts K., Hoemke L., Mohlberg H., Gilsbach J. & Zilles K. (2006). White matter fiber tracts of the human brain: Three-dimensional mapping at microscopic resolution, topography and intersubject variability. *NeuroImage* **29**(4):1092–1105.
- Burns J., Job D., Bastin M., Whalley H., Macgillivray T., Johnstone E. & Lawrie S. (2003). Structural disconnectivity in schizophrenia: a diffusion tensor magnetic resonance imaging study. *British Journal of Psychiatry* **182**:439–443.

- Callaghan P. (1991). *Principles of nuclear magnetic resonance microscopy*. Oxford University Press, UK.
- Callaghan P., Eccles C. & Xia Y. (1988). NMR microscopy of dynamic displacements: k-space and q-space imaging. *Journal of Physics E: Scientific Instruments* **21**(8):820–822.
- Campbell J., Siddiqi K., Rymar V., Sadikot A. & Pike G. (2005). Flow-based fiber tracking with diffusion tensor and q-ball data: validation and comparison to principal diffusion direction techniques. *NeuroImage* **27**(4):725–736.
- Catani M. & ffytche D. (2005). The rises and falls of disconnection syndromes. *Brain* **128**(10):2224–2239.
- Cherubini A., Luccichenti G., Péran P., Hagberg G., Barba C., Formisano R. & Sabatini U. (2007). Multimodal fMRI tractography in normal subjects and in clinically recovered traumatic brain injury patients. *NeuroImage* **34**(4):1331–1341.
- Chung S., Lu Y. & Henry R. (2006). Comparison of bootstrap approaches for estimation of uncertainties of DTI parameters. *NeuroImage* **33**(2):531–541.
- Ciccarelli O., Parker G., Toosy A., Wheeler-Kingshott C., Barker G., Boulby P., Miller D. & Thompson A. (2003a). From diffusion tractography to quantitative white matter tract measures: a reproducibility study. *NeuroImage* **18**(2):348–359.
- Ciccarelli O., Werring D., Barker G., Griffin C., Wheeler-Kingshott C., Miller D. & Thompson A. (2003b). A study of the mechanisms of normal-appearing white matter damage in multiple sclerosis using diffusion tensor imaging: Evidence of Wallerian degeneration. *Journal of Neurology* **250**(3):287–292.
- Clayden J., Bastin M. & Storkey A. (2006a). Automated assessment of tract similarity in group diffusion MRI data. In *Proceedings of the ISMRM 14th Scientific Meeting & Exhibition*, 2742. International Society for Magnetic Resonance in Medicine.
- Clayden J., Bastin M. & Storkey A. (2006b). Improved segmentation reproducibility in group tractography using a quantitative tract similarity measure. *NeuroImage* **33**(2):482–492.
- Clayden J., Marjoram D., Bastin M., Johnstone E. & Lawrie S. (2005). Towards an automated method for white matter integrity comparison between populations. In *Proceedings of the ESMRMB 22nd Annual Meeting*, 508. European Society for Magnetic Resonance in Medicine and Biology.
- Clayden J., Storkey A. & Bastin M. (2007a). A probabilistic model-based approach to consistent white matter tract segmentation. *IEEE Transactions on Medical Imaging* **26**(11):1555–1561.
- Clayden J., Zhang S., Correia S. & Laidlaw D. (2007b). Fine-grained comparison of anisotropy differences between groups of white matter tracts. In *Proceedings of the ISMRM-ESMRMB Joint Annual Meeting*, 82. International Society for Magnetic Resonance in Medicine.
- Concha L., Gross D. & Beaulieu C. (2006). Diffusion tensor imaging demonstrates the evolution of axonal degeneration after corpus callosotomy. In *Proceedings of the ISMRM 14th Scientific Meeting & Exhibition*, 274. International Society for Magnetic Resonance in Medicine.
- Conturo T., Lori N., Cull T., Akbudak E., Snyder A., Shimony J., McKinstry R., Burton H. & Raichle M. (1999). Tracking neuronal fiber pathways in the living human brain. *Proceedings of the National Academy of Sciences of the United States of America* **96**(18):10,422–10,427.
- Conturo T., McKinstry R., Akbudak E. & Robinson B. (1996). Encoding of anisotropic diffusion with tetrahedral gradients: a general mathematical diffusion formalism and experimental results. *Magnetic Resonance in Medicine* **35**(3):399–412.

- Cook P., Alexander D. & Parker G. (2004). Modelling noise-induced fibre-orientation error in diffusion-tensor MRI. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pp. 332–335. IEEE.
- Corouge I., Fletcher P., Joshi S., Gouttard S. & Gerig G. (2006). Fiber tract-oriented statistics for quantitative diffusion tensor MRI analysis. *Medical Image Analysis* **10**(5):786–798.
- Corouge I., Gouttard S. & Gerig G. (2004). *A statistical shape model of individual fiber tracts extracted from diffusion tensor MRI*, vol. 3217 of *Lecture Notes in Computer Science*. Springer-Verlag.
- de Boor C. (1978). *A practical guide to splines*. No. 27 in Applied Mathematical Sciences Series. Springer-Verlag.
- de Certaines J., Henriksen O., Spisni A., Cortsen M. & Ring P. (1993). In vivo measurements of proton relaxation times in human brain, liver, and skeletal muscle: a multicenter MRI study. *Magnetic Resonance Imaging* **11**(6):841–850.
- Deary I., Bastin M., Pattie A., Clayden J., Whalley L., Starr J. & Wardlaw J. (2006). White matter integrity and cognition in childhood and old age. *Neurology* **66**(4):505–512.
- Dempster A., Laird N. & Rubin D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**(1):1–38.
- Ding Z., Gore J. & Anderson A. (2003). Classification and quantification of neuronal fiber pathways using diffusion tensor MRI. *Magnetic Resonance in Medicine* **49**(4):716–721.
- Dirac P. (1958). *The principles of quantum mechanics*. Oxford University Press, UK, 4 edn.
- Edelstein W., Hutchison J., Johnson G. & Redpath T. (1980). Spin warp NMR imaging and applications to human whole-body imaging. *Physics in Medicine and Biology* **25**(4):751–756.
- Einstein A. (1905). [On the movement of small particles suspended in a stationary liquid demanded by the molecular-kinetic theory of heat]. *Annalen der Physik* **17**:549–560.
- Ellis C., Simmons A., Jones D., Bland J., Dawson J., Horsfield M., Williams S. & Leigh P. (1999). Diffusion tensor MRI assesses corticospinal tract damage in ALS. *Neurology* **53**(5):1051–1058.
- Evans A., Collins D., Mills S., Brown E., Kelly R. & Peters T. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. In *Nuclear Science Symposium and Medical Imaging Conference*, vol. 3, pp. 1813–1817. IEEE.
- ffytche D. & Catani M. (2005). Beyond localization: from hodology to function. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**(1456):767–779.
- Fick A. (1855). [On diffusion]. *Annalen der Physik und Chemie* **170**(1):59–86.
- Fick A. (1995). On liquid diffusion. *Journal of Membrane Science* **100**(1):33–38.
- Frank L. (2001). Anisotropy in high angular resolution diffusion-weighted MRI. *Magnetic Resonance in Medicine* **45**(6):935–939.
- Frank L. (2002). Characterization of anisotropy in high angular resolution diffusion-weighted MRI. *Magnetic Resonance in Medicine* **47**(6):1083–1099.
- Friman O., Farneback G. & Westin C.F. (2006). A Bayesian approach for stochastic white matter tractography. *IEEE Transactions on Medical Imaging* **25**(8):965–978.
- Friston K., Ashburner J., Frith C., Poline J., Heather J. & Frackowiak R. (1995). Spatial registration and normalization of images. *Human Brain Mapping* **3**(3):165–189.
- Friston K. & Frith C. (1995). Schizophrenia: a disconnection syndrome? *Clinical Neuroscience* **3**(2):89–97.

- Geman S. & Geman D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6):721–741.
- Geschwind N. (1965a). Disconnexion syndromes in animals and man: Part I. *Brain* **88**(2):237–294.
- Geschwind N. (1965b). Disconnexion syndromes in animals and man: Part II. *Brain* **88**(3):585–644.
- Good C., Johnsrude I., Ashburner J., Henson R., Friston K. & Frackowiak R. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage* **14**(1):21–36.
- Gray H. (1918). *Anatomy of the human body*. Lea & Febiger, Philadelphia, 20 edn.
- Guye M., Parker G., Symms M., Boulby P., Wheeler-Kingshott C., Salek-Haddadi A., Barker G. & Duncan J. (2003). Combined functional MRI and tractography to demonstrate the connectivity of the human primary motor cortex in vivo. *NeuroImage* **19**(4):1349–1360.
- Hahn E. (1950). Spin echoes. *Physical Review* **80**(4):580–594.
- Hasan K., Parker D. & Alexander A. (2001). Comparison of gradient encoding schemes for diffusion-tensor MRI. *Journal of Magnetic Resonance Imaging* **13**(5):769–780.
- Hastings W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1):97–109.
- Head D., Buckner R., Shimony J., Williams L., Akbudak E., Conturo T., McAvoy M., Morris J. & Snyder A. (2004). Differential vulnerability of anterior white matter in nondemented aging with minimal acceleration in dementia of the Alzheimer type: evidence from diffusion tensor imaging. *Cerebral Cortex* **14**(4):410–423.
- Healy D., Hendriks H. & Kim P. (1998). Spherical deconvolution. *Journal of Multivariate Analysis* **67**(1):1–22.
- Heiervang E., Behrens T., Mackay C., Robson M. & Johansen-Berg H. (2006). Between session reproducibility and between subject variability of diffusion MR and tractography measures. *NeuroImage* **33**(3):867–877.
- Holmes C., Hoge R., Collins D., Woods R., Toga A. & Evans A. (1998). Enhancement of MR images using registration for signal averaging. *Journal of Computer Assisted Tomography* **22**(2):324–333.
- Horsfield M. & Jones D. (2002). Applications of diffusion-weighted and diffusion tensor MRI to white matter diseases - a review. *NMR in Biomedicine* **15**(7-8):570–577.
- Hosey T., Williams G. & Ansorge R. (2005). Inference of multiple fiber orientations in high angular resolution diffusion imaging. *Magnetic Resonance in Medicine* **54**:1480–1489.
- Hu Q. & Nowinski W. (2003). A rapid algorithm for robust and automatic extraction of the midsagittal plane of the human cerebrum from neuroimages based on local symmetry and outlier removal. *NeuroImage* **20**(4):2153–2165.
- Huang H., Zhang J., Jiang H., Wakana S., Poetscher L., Miller M., van Zijl P., Hillis A., Wytik R. & Mori S. (2005). DTI tractography based parcellation of white matter: application to the mid-sagittal morphology of corpus callosum. *NeuroImage* **26**(1):195–205.
- Jansons K. & Alexander D. (2003). Persistent angular structure: new insights from diffusion magnetic resonance imaging data. *Inverse Problems* **19**(5):1031–1046.
- Jaynes E. (2003). *Probability theory: The logic of science*. Cambridge University Press, UK.

- Jenkinson M. & Smith S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis* 5(2):143–156.
- Johansen-Berg H., Behrens T., Robson M., Drobnjak I., Rushworth M., Brady J., Smith S., Higham D. & Matthews P. (2004). Changes in connectivity profiles define functionally distinct regions in human medial frontal cortex. *Proceedings of the National Academy of Sciences of the United States of America* 101(36):13,335–13,340.
- Johansen-Berg H., Behrens T., Sillery E., Ciccarelli O., Thompson A., Smith S. & Matthews P. (2005). Functional-anatomical validation and individual variation of diffusion tractography-based segmentation of the human thalamus. *Cerebral Cortex* 15(1):31–39.
- Jones D. (2003). Determining and visualizing uncertainty in estimates of fiber orientation from diffusion tensor MRI. *Magnetic Resonance in Medicine* 49(1):7–12.
- Jones D. (2004). The effect of gradient sampling schemes on measures derived from diffusion tensor MRI: a Monte Carlo study. *Magnetic Resonance in Medicine* 51(4):807–815.
- Jones D., Catani M., Pierpaoli C., Reeves S., Shergill S., O’Sullivan M., Golesworthy P., McGuire P., Horsfield M., Simmons A., Williams S. & Howard R. (2006). Age effects on diffusion tensor magnetic resonance imaging tractography measures of frontal cortex connections in schizophrenia. *Human Brain Mapping* 27(3):230–238.
- Jones D., Lythgoe D., Horsfield M., Simmons A., Williams S. & Markus H. (1999). Characterization of white matter damage in ischemic leukoaraiosis with diffusion tensor MRI. *Stroke* 30(2):393–397.
- Jones D. & Pierpaoli C. (2005). Confidence mapping in diffusion tensor magnetic resonance imaging tractography using a bootstrap approach. *Magnetic Resonance in Medicine* 53(5):1143–1149.
- Jones D., Symms M., Cercignani M. & Howard R. (2005a). The effect of filter size on VBM analyses of DT-MRI data. *NeuroImage* 26:546–554.
- Jones D., Travis A., Eden G., Pierpaoli C. & Basser P. (2005b). PASTA: Pointwise assessment of streamline tractography attributes. *Magnetic Resonance in Medicine* 53(6):1462–1467.
- Jönsson C. (1974). Electron diffraction at multiple slits. *American Journal of Physics* 42(1):4–11.
- Kaden E., Knosche T. & Anwander A. (2007). Parametric spherical deconvolution: Inferring anatomical connectivity using diffusion MR imaging. *NeuroImage* 37(2):474–488.
- Kanaan R., Shergill S., Barker G., Catani M., Ng V., Howard R., McGuire P. & Jones D. (2006). Tract-specific anisotropy measurements in diffusion tensor imaging. *Psychiatry Research: Neuroimaging* 146(1):73–82.
- Kingman J. & Taylor S. (1966). *Introduction to measure and probability*. Cambridge University Press, UK.
- Kinoshita M., Yamada K., Hashimoto N., Kato A., Izumoto S., Baba T., Maruno M., Nishimura T. & Yoshimine T. (2005). Fiber-tracking does not accurately estimate size of fiber bundle in pathological condition: initial neurosurgical experience using neuronavigation and subcortical white matter stimulation. *NeuroImage* 25(2):424–429.
- Klingberg T., Vaidya C., Gabrieli J., Moseley M. & Hedehus M. (1999). Myelination and organization of the frontal white matter in children: a diffusion tensor MRI study. *Neuroreport* 10(13):2817–2821.
- Knight R., Ordidge R., Helpert J., Chopp M., Rodolosi L. & Peck D. (1991). Temporal evolution of ischemic damage in rat brain measured by proton nuclear magnetic resonance imaging. *Stroke* 22(6):802–808.

- Kochunov P., Thompson P., Lancaster J., Bartzokis G., Smith S., Coyle T., Royall D., Laird A. & Fox P. (2007). Relationship between white matter fractional anisotropy and other indices of cerebral health in normal aging: tract-based spatial statistics study of aging. *NeuroImage* **35**(2):478–487.
- Kubicki M., Westin C.F., Nestor P., Wible C., Frumin M., Maier S., Kikinis R., Jolesz F., McCarley R. & Shenton M. (2003). Cingulate fasciculus integrity disruption in schizophrenia: a magnetic resonance diffusion tensor imaging study. *Biological Psychiatry* **54**(11):1171–1180.
- Lanzer P., Botvinick E., Schiller N., Crooks L., Arakawa M., Kaufman L., Davis P., Herfkens R., Lipton M. & Higgins C. (1984). Cardiac imaging using gated magnetic resonance. *Radiology* **150**(1):121–127.
- Lazar M. & Alexander A. (2002). White matter tractography using RandomVector (RAVE) perturbation. In *Proceedings of the ISMRM 10th Scientific Meeting & Exhibition*, p. 539. International Society for Magnetic Resonance in Medicine.
- Lazar M. & Alexander A. (2005). Bootstrap white matter tractography (BOOT-TRAC). *NeuroImage* **24**(2):524–532.
- Le Bihan D., Breton E., Lallemand D., Grenier P., Cabanis E. & Laval-Jeantet M. (1986). MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology* **161**(2):401–407.
- Lim K. & Helpert J. (2002). Neuropsychiatric applications of DTI - a review. *NMR in Biomedicine* **15**(7-8):587–593.
- Lori N., Conturo T. & Le Bihan D. (2003). Definition of displacement probability and diffusion time in q-space magnetic resonance measurements that use finite-duration diffusion-encoding gradients. *Journal of Magnetic Resonance* **165**(2):185–195.
- MacKay D. (1995). Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* **6**(3):469–505.
- MacKay D. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press, UK.
- Maddah M., Mewes A., Haker S., Grimson W. & Warfield S. (2005). Automated atlas-based clustering of white matter fiber tracts from DTMRI. In *Medical Image Computing and Computer-Assisted Intervention*, vol. 3749 of *Lecture Notes in Computer Science*, pp. 188–195. Springer-Verlag.
- Makris N., Kennedy D., McInerney S., Sorensen A., Wang R., Caviness V. & Pandya D. (2005). Segmentation of subcomponents within the superior longitudinal fascicle in humans: a quantitative, in vivo, DT-MRI study. *Cerebral Cortex* **15**(6):854–869.
- Makris N., Papadimitriou G., van der Kouwe A., Kennedy D., Hodge S., Dale A., Benner T., Wald L., Wu O., Tuch D., Caviness V., Moore T., Killiany R., Moss M. & Rosene D. (2007). Frontal connections and cognitive changes in normal aging rhesus monkeys: a DTI study. *Neurobiology of Aging* **28**(10):1556–1567.
- Mansfield P. (1977). Multi-planar image formation using NMR spin echoes. *Journal of Physics C: Solid State Physics* **10**(3):L55–L58.
- Mattiello J., Basser P. & Le Bihan D. (1997). The b matrix in diffusion tensor echo-planar imaging. *Magnetic Resonance in Medicine* **37**(2):292–300.
- McIntosh A., Job D., Moorhead T., Harrison L., Forrester K., Lawrie S. & Johnstone E. (2004). Voxel-based morphometry of patients with schizophrenia or bipolar disorder and their unaffected relatives. *Biological Psychiatry* **56**:544–552.

- Metropolis N., Rosenbluth A., Rosenbluth M., Teller A. & Teller E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**(6):1087–1092.
- Mori S., Crain B., Chacko V. & van Zijl P. (1999). Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of Neurology* **45**(2):265–269.
- Mori S. & van Zijl P. (2002). Fiber tracking: principles and strategies - a technical review. *NMR in Biomedicine* **15**:468–480.
- Mori S., Wakana S., Nagae-Poetscher L. & van Zijl P. (2005). *MRI atlas of human white matter*. Elsevier, Amsterdam, The Netherlands.
- Morris P. (1986). *Nuclear magnetic resonance imaging in medicine and biology*. Oxford University Press, UK.
- Moseley M. (2002). Diffusion tensor imaging and aging - a review. *NMR in Biomedicine* **15**(7-8):553–560.
- Moseley M., Kucharczyk J., Asgari H. & Norman D. (1991). Anisotropy in diffusion-weighted MRI. *Magnetic Resonance in Medicine* **19**(2):321–326.
- Muñoz Maniega S., Bastin M., Armitage P., Farrall A., Carpenter T., Hand P., Cvorov V., Rivers C. & Wardlaw J. (2004). Temporal evolution of water diffusion parameters is different in grey and white matter in human ischaemic stroke. *Journal of Neurology Neurosurgery and Psychiatry* **75**(12):1714–1718.
- Neal R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. *Tech. Rep. CRG-TR-93-1*, University of Toronto.
- Nieuwenhuys R. (1996). *The human central nervous system: a synopsis and atlas*. Springer, New York, 3 edn.
- O'Donnell L. & Westin C.F. (2005). White matter tract clustering and correspondence in populations. In *Medical Image Computing and Computer-Assisted Intervention*, vol. 3749 of *Lecture Notes in Computer Science*, pp. 140–147. Springer-Verlag.
- O'Sullivan M., Jones D., Summers P., Morris R., Williams S. & Markus H. (2001). Evidence for cortical "disconnection" as a mechanism of age-related cognitive decline. *Neurology* **57**(4):632–638.
- Pajevic S. & Basser P. (2003). Parametric and non-parametric statistical analysis of DT-MRI data. *Journal of Magnetic Resonance* **161**(1):1–14.
- Papadakis N., Xing D., Houston G., Smith J., Smith M., James M., Parsons A., Huang C., Hall L. & Carpenter T. (1999). A study of rotationally invariant and symmetric indices of diffusion anisotropy. *Magnetic Resonance Imaging* **17**(6):881–892.
- Park H.J., Westin C.F., Kubicki M., Maier S., Niznikiewicz M., Baer A., Frumin M., Kikinis R., Jolesz F., McCarley R. & Shenton M. (2004). White matter hemisphere asymmetries in healthy subjects and in schizophrenia: a diffusion tensor MRI study. *NeuroImage* **23**:213–223.
- Parker G., Haroon H. & Wheeler-Kingshott C. (2003). A framework for a streamline-based probabilistic index of connectivity (PICO) using a structural interpretation of MRI diffusion measurements. *Journal of Magnetic Resonance Imaging* **18**(2):242–254.
- Parker G., Stephan K., Barker G., Rowe J., MacManus D., Wheeler-Kingshott C., Ciccarelli O., Passingham R., Spinks R., Lemon R. & Turner R. (2002a). Initial demonstration of in vivo tracing of axonal projections in the macaque brain and comparison with the human brain using diffusion tensor imaging and fast marching tractography. *NeuroImage* **15**(4):797–809.
- Parker G., Wheeler-Kingshott C. & Barker G. (2002b). Estimating distributed anatomical connectivity using fast marching methods and diffusion tensor imaging. *IEEE Transactions on Medical Imaging* **21**(5):505–512.



- Perrin M., Poupon C., Rieul B., Leroux P., Constantinesco A., Mangin J. & Le Bihan D. (2005). Validation of q-ball imaging with a diffusion fibre-crossing phantom on a clinical scanner. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**(1457):881–891.
- Pfefferbaum A., Sullivan E., Hedehus M., Lim K., Adalsteinsson E. & Moseley M. (2000). Age-related decline in brain white matter anisotropy measured with spatially corrected echo-planar diffusion tensor imaging. *Magnetic Resonance in Medicine* **44**(2):259–268.
- Pierpaoli C. & Basser P. (1996). Toward a quantitative assessment of diffusion anisotropy. *Magnetic Resonance in Medicine* **36**(6):893–906.
- Pipe J. (1999). Motion correction with PROPELLER MRI: application to head motion and free-breathing cardiac imaging. *Magnetic Resonance in Medicine* **42**(5):963–969.
- Poupon C., Clark C., Frouin V., Régis J., Bloch I., Le Bihan D. & Mangin J. (2000). Regularization of diffusion-based direction maps for the tracking of brain white matter fascicles. *NeuroImage* **12**(2):184–195.
- Price G., Cercignani M., Parker G., Altmann D., Barnes T., Barker G., Joyce E. & Ron M. (2007). Abnormal brain connectivity in first-episode psychosis: a diffusion MRI tractography study of the corpus callosum. *NeuroImage* **35**(2):458–466.
- Pusey E., Lufkin R., Brown R., Solomon M., Stark D., Tarr R. & Hanafey W. (1986). Magnetic resonance imaging artifacts: mechanism and clinical significance. *Radiographics* **6**(5):891–911.
- R Development Core Team (2007). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Reese T., Heid O., Weisskoff R. & Wedeen V. (2003). Reduction of eddy-current-induced distortion in diffusion MRI using a twice-refocused spin echo. *Magnetic Resonance in Medicine* **49**(1):177–182.
- Riley K., Hobson M. & Bence S. (2002). *Mathematical methods for physics and engineering*. Cambridge University Press, UK, 2 edn.
- Rouw R. & Scholte H. (2007). Increased structural connectivity in grapheme-color synesthesia. *Nature Neuroscience* **10**(6):792–797.
- Sakaie K. & Lowe M. (2007). An objective method for regularization of fiber orientation distributions derived from diffusion-weighted MRI. *NeuroImage* **34**(1):169–176.
- Schmahmann J., Pandya D., Wang R., Dai G., D’Arceuil H., de Crespigny A. & Wedeen V. (2007). Association fibre pathways of the brain: parallel observations from diffusion spectrum imaging and autoradiography. *Brain* **130**(3):630–653.
- Sebastian T., Klein P. & Kimia B. (2003). On aligning curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(1):116–125.
- Skare S., Hedehus M., Moseley M. & Li T. (2000). Condition number as a measure of noise performance of diffusion tensor data acquisition schemes with MRI. *Journal of Magnetic Resonance* **147**:340–352.
- Smith S., Jenkinson M., Johansen-Berg H., Rueckert D., Nichols T., Mackay C., Watkins K., Ciccarelli O., Cader M., Matthews P. & Behrens T. (2006). Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage* **31**(4):1487–1505.
- Smith S., Jenkinson M., Woolrich M., Beckmann C., Behrens T., Johansen-Berg H., Bannister P., De Luca M., Drobnjak I., Flitney D., Niazy R., Saunders J., Vickers J., Zhang Y., De Stefano N., Brady J. & Matthews P. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23 Suppl 1**:S208–S219.

- Staempfli P., Jaermann T., Crelier G., Kollias S., Valavanis A. & Boesiger P. (2006). Resolving fiber crossing using advanced fast marching tractography based on diffusion tensor imaging. *NeuroImage* **30**(1):110–120.
- Staempfli P., Reischauer C., Jaermann T., Valavanis A., Kollias S. & Boesiger P. (2008). Combining fMRI and DTI: A framework for exploring the limits of fMRI-guided DTI fiber tracking and for verifying DTI-based fiber tractography results. *NeuroImage* **39**(1):119–126.
- Stejskal E. (1965). Use of spin echoes in a pulsed magnetic-field gradient to study anisotropic, restricted diffusion and flow. *Journal of Chemical Physics* **43**(10):3597–3603.
- Stejskal E. & Tanner J. (1965). Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *Journal of Chemical Physics* **42**(1):288–292.
- Sullivan E. & Pfefferbaum A. (2006). Diffusion tensor imaging and aging. *Neuroscience & Biobehavioral Reviews* **30**(6):749–761.
- Tournier J.D., Calamante F. & Connelly A. (2007). Robust determination of the fibre orientation distribution in diffusion MRI: Non-negativity constrained super-resolved spherical deconvolution. *NeuroImage* **35**(4):1459–1472.
- Tournier J.D., Calamante F., Gadian D. & Connelly A. (2004). Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *NeuroImage* **23**(3):1176–1185.
- Tuch D. (2004). Q-ball imaging. *Magnetic Resonance in Medicine* **52**(6):1358–1372.
- Tuch D., Reese T., Wiegell M., Makris N., Belliveau J. & Wedeen V. (2002). High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. *Magnetic Resonance in Medicine* **48**(4):577–582.
- Tuch D., Reese T., Wiegell M. & Wedeen V. (2003). Diffusion MRI of complex neural architecture. *Neuron* **40**(5):885–895.
- Volkau I., Bhanu Prakash K., Ananthasubramaniam A., Aziz A. & Nowinski W. (2006). Extraction of the midsagittal plane from morphological neuroimages using the Kullback-Leibler's measure. *Medical Image Analysis* **10**(6):863–874.
- Wedeen V., Hagmann P., Tseng W.Y., Reese T. & Weisskoff R. (2005). Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. *Magnetic Resonance in Medicine* **54**(6):1377–1386.
- Whitcher B., Tuch D., Wisco J., Sorensen A. & Wang L. (2007). Using the wild bootstrap to quantify uncertainty in DTI. *Human Brain Mapping* **in press**.
- Woods R., Grafton S., Watson J., Sicotte N. & Mazziotta J. (1998). Automated image registration: II. Intersubject validation of linear and nonlinear models. *Journal of Computer Assisted Tomography* **22**(1):153–165.
- Wright I., McGuire P., Poline J., Travers J., Murray R., Frith C., Frackowiak R. & Friston K. (1995). A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. *NeuroImage* **2**(4):244–252.
- Xue R., van Zijl P., Crain B., Solaiyappan M. & Mori S. (1999). In vivo three-dimensional reconstruction of rat brain axonal projections by diffusion tensor imaging. *Magnetic Resonance in Medicine* **42**(6):1123–1127.
- Zhang S., Demiralp Ç. & Laidlaw D. (2003). Visualizing diffusion tensor MR images using streamtubes and streamsurfaces. *IEEE Transactions on Visualization and Computer Graphics* **9**(4):454–462.

---

# Colophon

THIS THESIS was typeset using pdf $\text{\TeX}$  on Mac os x, based on the  $\text{\TeX}$  typesetting system by Donald Knuth. The text is set in the Palatino typeface, while titles and captions are set in Optima, both of which were designed by Hermann Zapf in the 1940s–1950s. The latter font is made available to  $\text{\LaTeX}$  thanks to the `gtamacoptima` package by Thomas A. Schmitz and Adam Lindsay, distributed as part of the `gwTeX` distribution of  $\text{\TeX}$  for the Mac by Gerben Wierda. All mathematical nomenclature is also based on the Palatino typeface, and credit for this extension is due to Young Ryu’s `pxfonts` package.  $\text{\TeX}$  code for customised chapter headings was based on a template by Vincent Zoonekynd. The  $\text{\LaTeX}$  packages `algorithmic`, `algorithm`, `ambsy`, `amsmath`, `caption`, `color`, `fontenc`, `graphicx`, `inputenc`, `latexsym`, `lettrine`, `mathrsfs`, `natbib`, `pifont`, `relsize`, `subfigure`, `supertabular` and `xspace` were also used to create the document, and the author is grateful to all those who contributed to them. The document class file was written by Martin Reddy, Ellen Foster and Tiejun Ma at the University of Edinburgh.

Almost all software development was based on the `R` language and environment (R Development Core Team, 2007). Source editing was made considerably easier thanks to the versatile TextMate editor, by Allan Odgaard. Most scientific figures were created with either `R` or Mathematica—the latter is developed by Wolfram Research, Inc. Compositing tract images were created with the help of the open source ImageMagick library, developed and maintained by ImageMagick Studio LLC. Diagramming and layout was done with the OmniGraffle package, by Omni Development, Inc.