8

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

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

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.

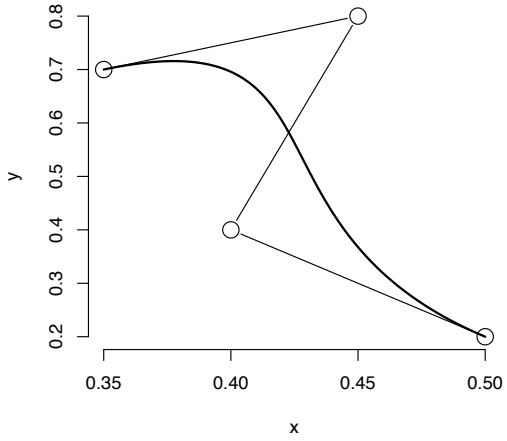

Figure 8.1: Two-dimensional cubic Bézier curve (thick line) with control points indicated with open circles.

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

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

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

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

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.

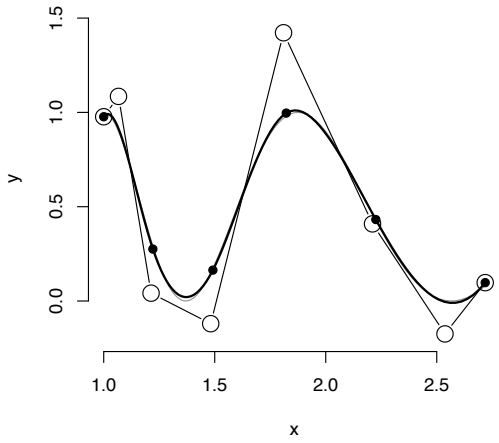

Figure 8.2: Cubic B-spline approximation to the 2-D parametric function $(e^t, \cos^2 5t)$ over the interval $[0,1]$, using four internal knots. The thick black line is the spline curve, filled circles represent knot points and open circles control points. The true function is shown in grey, but it is 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 ξ-quantiles of the individual streamline lengths, where ξ 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 unterminated streamlines. The resultant set of median points is a single line tract representation $r = (\mathbf{x}_i)$, where $i \in \{-\tilde{N}_1, -\tilde{N}_1 + 1, \ldots, \tilde{N}_2 - 1, \tilde{N}_2\}$ 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 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 unterminated 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

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

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, $\boldsymbol{\rho}_i$, at each point, i, on the median line are used to calculate the residual standard error, according to

$$\mathbf{E}_\kappa = \sqrt{\frac{\sum_i {\boldsymbol{\rho}_i}^2}{\tilde{N}_1 + \tilde{N}_2 - \kappa - 3}} \,. \tag{8.5}$$

(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, η. 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.

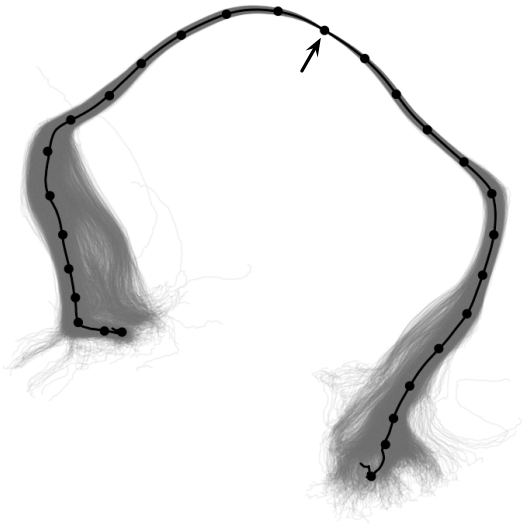

Figure 8.3: Graphical representation of a full set of probabilistic streamlines representing the corpus callosum splenium (grey, shown at 10% opacity), the median line and B-spline knot points (black), here projected into a plane normal to the superior–inferior (z) axis. $\xi = 0.99$. The seed point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

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 $\mathbf{v}^i_u$, 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

$$c^i_u = \cos\theta^i_u = \frac{\mathbf{v}^i_u \cdot \mathbf{w}^i_u}{\|\mathbf{v}^i_u\|\|\mathbf{w}^i_u\|} \tag{8.6}$$

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

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

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

$$s^i_u = \cos\phi^i_u = \frac{\mathbf{v}^i_u \cdot \mathbf{v}^*_u}{\|\mathbf{v}^i_u\|\|\mathbf{v}^*_u\|}\,, \tag{8.7}$$

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

Fig. 8.4 illustrates, in two dimensions, the continuity angles, θ_u, and the similarity angles, ϕ_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: $\mathbf{d}^i = (L^i_1, L^i_2, \mathbf{c}^i, \mathbf{s}^i)$, where $\mathbf{c}^i = (c^i_u)$ and $\mathbf{s}^i = (s^i_u)$. 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

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

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,

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

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

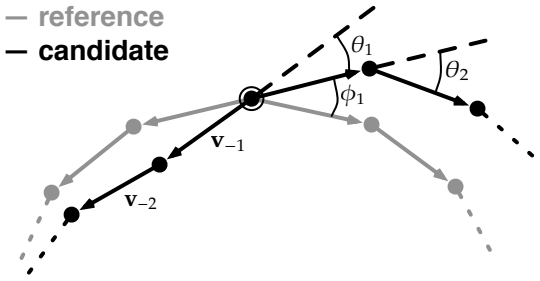

Figure 8.4: Illustration of the different angles relevant to our model. Filled circles here represent successive knot points in the reference and candidate tracts. The ringed knot is the seed point, which is common to the two tracts.

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

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

where $\Gamma(\cdot)$ is the gamma function, and α and β 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

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

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

To find maximum likelihood estimates for α and ε 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, ζ_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, η, 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

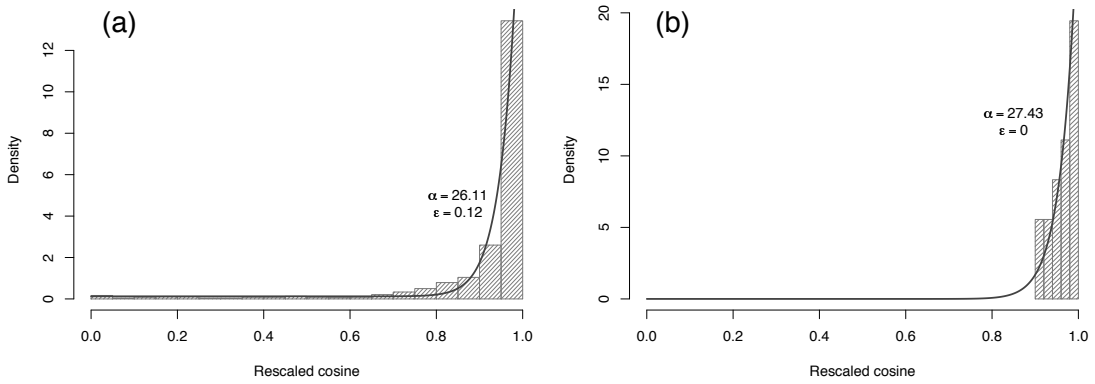

Figure 8.5: Histograms of rescaled **(a)** continuity cosines ($n = 962$) and **(b)** similarity cosines ($u = 7$, $n = 18$) from the splenium training data. The appropriate density functions from the model are overlaid.

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 ε reflects this. In (b), however, there are no cosine data below 0.9, and so the ε parameter has shrunk to zero. In fact, all of the similarity cosine distributions had $\varepsilon = 0$, although the α 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 7×7×7 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.

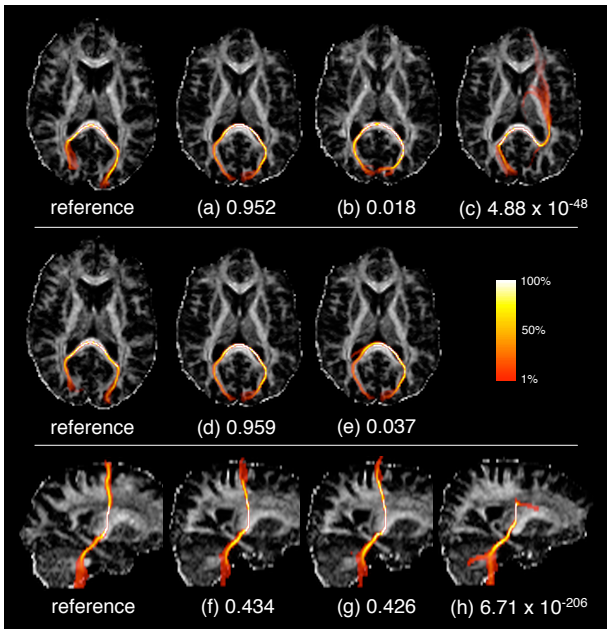

Figure 8.6: The two most likely matches to the original (top row) and the alternative (middle row) splenium reference tract, shown in axial projection with their associated matching probabilities. The tract generated from the neighbourhood centre point is shown with its matching probability **(c)**, 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,

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

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

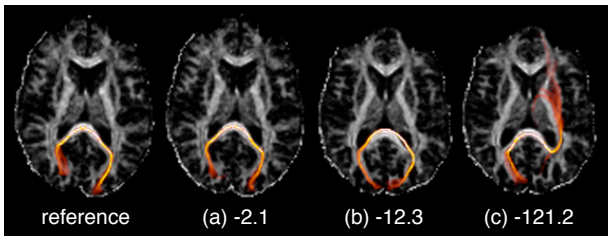

Figure 8.7: Log-ratios between matching likelihoods of the tracts shown and the matching likelihood of the reference tract. The reference tract has a log-ratio of zero by definition; **(a)** is the alternative reference tract; **(b)** is the best match in the novel candidate set; and **(c)** is the tract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

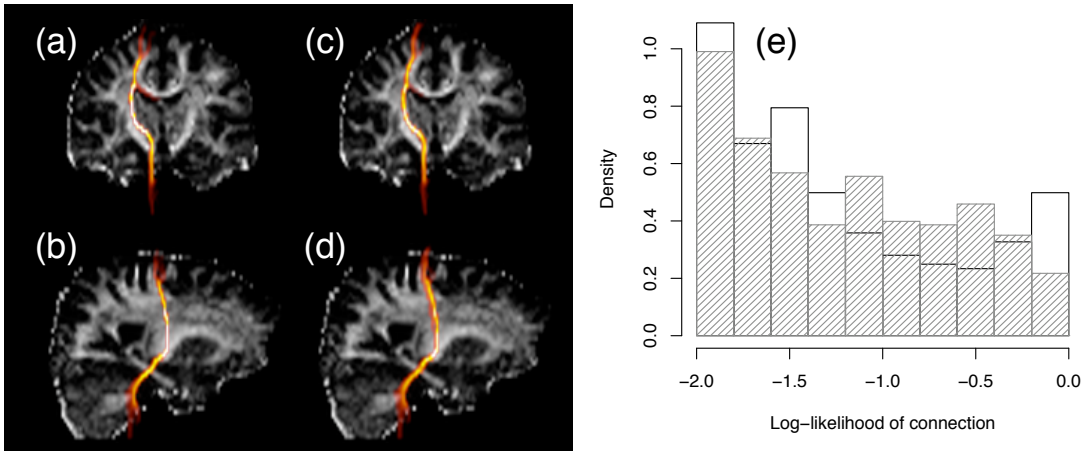

Figure 8.8: Maximum likelihood corticospinal tract images in coronal **(a)** and sagittal **(b)** projections, and equivalent images for the weighted average tract data **(c,d)**, all thresholded at the 1% level. Histograms of connection log-likelihoods for the maximum likelihood (plain) and weighted 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-ROI constraint could be applied to ameliorate problems with branches in the tracts if they proved significant, although we would advocate the avoidance of ROI 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

$$\sum_i P(z^i = 1) = 1 \,. \tag{8.12}$$

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, $\hat{\boldsymbol{\omega}}$, the E-step of the algorithm would then involve calculating the posteriors

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

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

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

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

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

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 γ, 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,

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

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 α_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 α 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 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 α parameter to regularise the likelihood function, and then take the maximum *a posteriori* estimate

$$\hat{\alpha} = \arg\max_{\alpha}\left\{\ln P(\alpha\,|\,\mathbf{x})\right\} = \arg\max_{\alpha}\left\{\sum_i \ln P(x^i\,|\,\alpha) + \ln P(\alpha)\right\}.$$

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

$$\hat{\alpha}_u = \frac{-2\sum_{i>0} P(z^i = 1\,|\,D)}{\sum_{i>0} P(z^i = 1\,|\,D)\ln x_u^i - \lambda}. \tag{8.20}$$

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

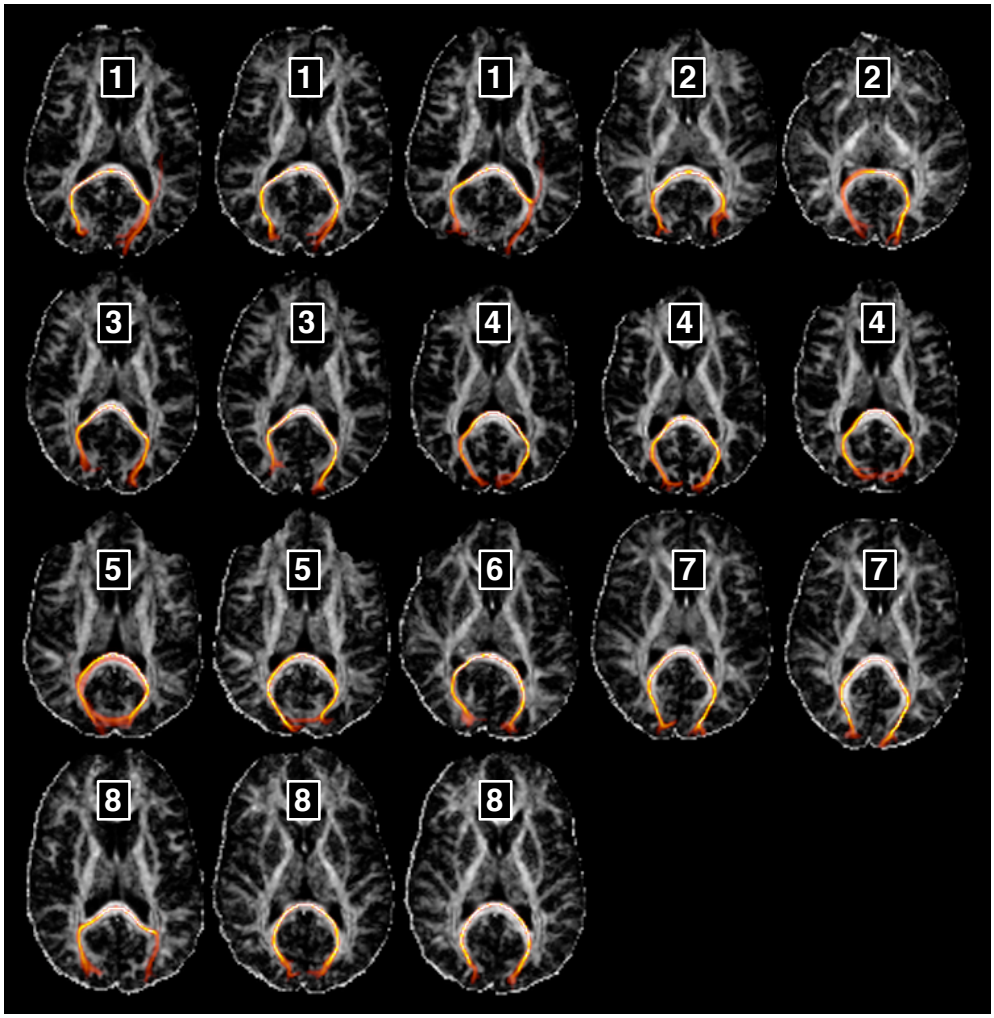

Figure 8.9: Best matching splenium tracts from a full data set of 18 scans, selected using the EM approach to neighbourhood tractography and thresholded at 1%. The numbers indicate the subject number from which each scan was taken. The reference tract was taken from another scan of subject 3.

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 agèd 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.