Probabilistic Clustering of White Matter Fibre Bundles using Regression Mixtures

Nagulan Ratnarajah[1]

Andy Simmons[2]

Ali Hojjatoleslami[1]

[1] Neurosciences and Medical Image Computing, University of Kent, U.K.

[2] Institute of Psychiatry, Kings College London, U.K.

Abstract

We present a novel approach for probabilistic clustering of white matter fibre pathways using curve-based regression mixture modelling techniques in 3D curve space. The clustering algorithm is based on a principled method for probabilistic modelling of a set of fibre trajectories as individual sequences of points generated from a finite mixture model consisting of multivariate polynomial regression model components. Unsupervised learning is carried out using maximum likelihood principles. Specifically, conditional mixture is used together with an EM algorithm to estimate cluster membership. The result of clustering is a probabilistic assignment of fibre trajectories to each cluster and an estimate of cluster parameters. A statistical model is calculated for each clustered fibre bundle using fitted parameters of the probabilistic clustering. We illustrate the potential of our clustering approach on synthetic and real data.

1 Introduction

White matter (WM) fibre clustering is becoming an important field of clinical neuroscience research since it facilitates insights about anatomical structures in health and disease, allows clear visualizations of fibre tracts and enables the calculation of relevant statistics across subjects. A number of algorithms have been developed for clustering and labelling WM fibre bundles in DTI. Deterministic clustering algorithms [1-3] assign each trajectory to only one cluster, which may lead to biased estimators of cluster parameters if the clusters overlap. Probabilistic clustering algorithms [4], on the contrary, deal with the inherent uncertainty in assigning the trajectories to clusters. Quantitative parameters can be estimated by a weighted average over cluster members and thus more robust results may be obtained, which are less sensitive to the presence of outliers. Maddah et al. [4] proposed a probabilistic approach using a gamma mixture model and a distance map. This method assumes that the number of clusters is known and the approach requires manual user initialisation of the cluster centres. A problem for this approach was establishing correspondence between points.

In this paper, we propose a new geometrical framework to automatically cluster WM fibres into biologically meaningful neuro-tracts probabilistically. Specifically we use mixtures of polynomial regression models as the basis of clustering. Multivariate clustering technique is used to describe the three dimensional propagations of the fibre trajectories which vary in length. We use a conditional mixture approach as it naturally allows for curves of variable length with unique measurement intervals and missing observations. Polynomial fits also take advantage of smoothness information present in the data. A regression model for each fibre bundle is constructed after performing probabilistic clustering. The probabilistic clustering algorithm is also capable of handling outliers in a principled way.

2 Probabilistic Model for White Matter Trajectories

Basic Definitions: Let V be a set of M 3-D fibre trajectories, where each trajectory v_i is an $n_i \times 3$ matrix containing a sequence of n_i 3-D points (x,y,z) in $\Re$. The associated $n_i \times 1$ vector u_i of ordered points from 0 to $n_i - 1$ correspond to points of v_i and set $U = \{u_1, u_2, \ldots, u_M\}$. In the standard mixture model framework probability density function (PDF) for a d-dimensional vector v, is modelled as a function of model parameters φ, by the mixture density

$$p(v|\varphi) = \textstyle\sum_k^K \alpha_k p_k(v|\theta_k), \qquad (1)$$

in which $\varphi = \{\alpha_k; \theta_k, k = 1 \ldots K\}, \alpha_k(\sum \alpha_k = 1)$ is the k-th component weight and p_k is the k-th component density with parameter vector θ_k.

In this manner a finite mixture model is a PDF composed of a weighted average of component density functions. Each trajectory v_i is generated by one of the components, but the identity of the generating component is not observed. The parameters of each density component $p_k(v|\theta_k)$, as well as the corresponding weights α_k, can be estimated from the data using the EM algorithm. The estimated component models, $p_k(v|\theta_k)$ are interpreted as K clusters, where each cluster is defined by a PDF. The set of trajectories is clustered to a number of subsets by assigning a membership probability, w_{ik}, to each trajectory, v_i, to denote its membership of the kth cluster. The number of clusters, K, is defined by the user. Finally, each trajectory v_i is assigned to the cluster k with the highest membership probability.

Model Definition: We model the X directional position (similarly Y and Z) with a p-th order multivariate polynomial regression model in which the order u_i is the independent variable, which is assumed with an additive Gaussian error term. The three regression equations can be defined succinctly in terms of the matrix v_i. The form of the regression equation for v_i is

$$v_i = U_i\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0,\Sigma) \qquad (2)$$

where U_i is the standard $n_i \times (p+1)$ Vandermonde regression matrix associated with vector u_i, β is a $(p+1) \times 3$ matrix of regression coefficients for X, Y, and Z direction and ε_i is an $n_i \times 3$ zero-mean matrix multivariate normal error term with a covariance matrix Σ. For simplicity, we assume that $\Sigma = diag(\sigma_x^2, \sigma_y^2, \sigma_z^2)$, so that the X, Y, and Z measurement noise terms are treated as conditionally independent given the model.

The conditional density for the ith trajectory f is a multivariate Gaussian with matrix mean $U_i\beta$ and covariance matrix Σ. The parameter set $\theta = \{\beta, \Sigma\}$.

$$p(v_i|u_i,\theta) = f(v_i|U_i\beta,\Sigma) = (2\Pi)^{-n_i}|\Sigma|^{-\frac{n_i}{2}} exp\{-\tfrac{1}{2}tr[(v_i - U_i\beta)\Sigma^{-1}(v_i - U_i\beta)^{'}]\} \qquad (3)$$

We can derive regression mixtures for the trajectories by a substitution of Eq (1) with the conditional regression density components $p_k(v|u,\theta_k)$, as defined in Eq (3).

$$p(v_i|u_i,\varphi) = \textstyle\sum_k^K \alpha_k f_k(v_i|U_i\beta_k, \Sigma_k) \qquad (4)$$

Note that in this model each fibre trajectory is assumed to be generated by one of K different regression models. Each model has its own shape parameters $\theta_k = \{\beta_k, \Sigma_k\}$.

The full probability density V given U, $p(V|U,\varphi)$, is also known as the conditional likelihood of the parameter φ given the data set both V and U to be written as

$$L(\varphi|V,U) = p(V|U,\varphi) = \prod_i^N \sum_k^K \alpha_k f_k(v_i|U_i\beta_k,\Sigma_k) \qquad (5)$$

The model can handle trajectories of variable length in a natural fashion, since the likelihood equation (Eq (5)) does not require the number of data points. The product form in Eq (5) follows from assuming conditional independence of v_i's, given both u_i's and the mixture model, since the fibre trajectories do not influence each other.

EM Algorithm for Mixture of Regression: In the E-step, we estimate the hidden cluster memberships by forming the ratio of the likelihood of trajectory v_i under cluster k, to the sum-total likelihood of trajectory v_i under all clusters:

$$w_{ik} = \frac{\alpha_k f_k(v_i|U_i\beta_k,\Sigma_k)}{\sum_{j=1}^K \alpha_j f_j(v_i|U_i\beta_j,\Sigma_j)} \qquad (6)$$

These w_{ik} give the probabilities that the ith trajectory was generated from cluster k.

In the M-step, the expected cluster memberships from the E-step are used to form the weighted log-likelihood function: $L(\varphi|V,U) = \sum_i \sum_k w_{ik} \; log \alpha_k f_k(v_i|U_i\beta_k,\Sigma_k)$ (7)
The membership probabilities weight the contribution that the kth density component adds to the overall likelihood. The weighted log-likelihood is then maximized with respect to the parameter set φ.

Let $w_{ik} = w_{ik} I_{n_i}$, where I_{n_i} is an identity vector, and let $W_k = diag(w'_{1k}, w'_{2k}, \ldots, w'_{nk})$ be an $N \times N$ diagonal matrix, where $N = \Sigma_i^M n_i$. Then, we use W_k to calculate the mixture parameters.

$$\hat{\beta}_k = (U' W_k U)^{-1} U' W_k V, \;\; \hat{\Sigma}_k = \frac{(V - U\hat{\beta}_k)' W_k (V - U\hat{\beta}_k)}{\sum_i^N w_{ik}}, \;\; \text{and } \hat{\alpha}_k = \frac{1}{N}\sum_i w_{ik} \;\; \text{for } k = 1,\ldots,K \qquad (8)$$

where V is an $N \times 3$ matrix containing all the v_i measurements, one trajectory after another, and U is an $N \times (p+1)$ matrix corresponding to Y values.

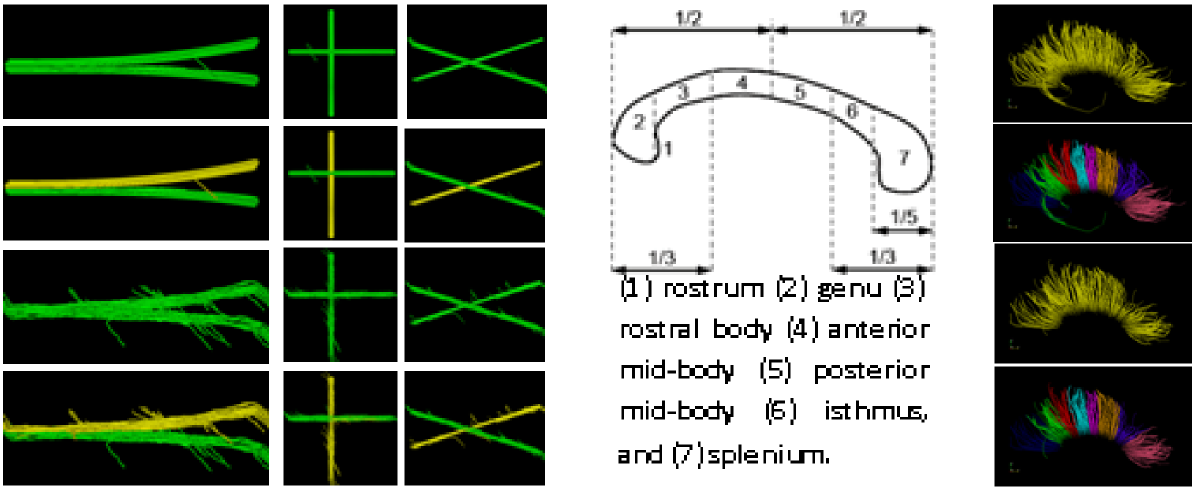
3 Methods

Synthetic Data: We have used PISTE [http://cubric.psych.cf.ac.uk/commondti] synthetic data set (diffusion encoding directions = 30, b-value = 1000 s/mm^2 and voxel resolution: $1 \times 1 \times 1$ mm^3) to demonstrate some of the basic features of our clustering algorithm, specifically, its ability to cluster a 3D data set into multiple bundles accurately. Here we consider three example noise free and noisy (SNR=15) data sets: a branching fibre, two orthogonally crossing fibres and two straight crossing fibres. For the 3D tract reconstruction, the single-tensor and two-tensor 4th order Runge-Kutta method were used for branching data and two fibre crossing data respectively.

***In Vivo* Data:** 1.5 T DW data were acquired from four healthy adults with an image matrix of 128x128, 60 slice locations covering the whole brain, $1.875 \times 1.875 \times 2.0$ mm^3 spatial resolution, b = 700 s/mm^2 and 41 diffusion directions. To correct for eddy currents and motion, each DW volume was registered to the non-DW volume of the first subject.

Corpus Callosum Clustering: Subdividing the corpus callosum (CC) into anatomically distinct regions is not well defined but is of much importance, especially in studying normal development and in understanding psychiatric and neurodegenerative disorders. Witelson [5] proposed a schematic for seven subdivisions of the CC as shown in Figure 1(ii). We further divide the splenium into its upper and lower parts to give a finer model.

The ROIs for the CC were outlined by an expert based on information from FA maps for all four subjects. Fibre trajectories were reconstructed using the 4th order Runge-Kutta method for the four subjects and were normalized to a common template (128x128x60 matrix size and voxel size 1x1x1 unit). The CC tracts were then clustered into K=8 subdivisions.


1/2
1/2
1
2
3
4
5
6
7
1/5
1/3
1/3
(1) rostrum (2) genu (3) rostral body (4) anterior mid-body (5) posterior mid-body (6) isthmus, and (7) splenium.

(i) (ii) (iii)

Figure 1: (i) Synthetic data, clustered trajectories, noisy data and noisy data clustered trajectories in rows (ii) A schematic of Witelson corpus callosum subdivisions [5] based on the midsaggital slice (iii) Clustering of the CC from the first two subjects viewed from a sagittal orientation: the original fibre tracts (yellow) and clustered into bundles.

The choice to use third-order polynomials for the regression models as opposed to other order polynomials was made for two reasons: (a) visual inspection supports this as a sufficient choice and (b) cross-validation also confirms third-order as the optimal choice in this case. We modelled the X position with a cubic polynomial regression model in which u is the independent variable, $x = \beta_3 u^3 + \beta_2 u^2 + \beta_1 u + \beta_0$, and likewise for the Y and Z directions.

4 Results and Discussion

Synthetic Data: The Synthetic data results demonstrate the clustering algorithm's ability to accurately separate fibre tracts into meaningful bundles. In our component regression models for the synthetic data a cubic polynomial was used (K=2). This choice is based on the visual inspection of fitted-versus-actual trajectory data. The noise-free synthetic data results in complicated fibre tract structures demonstrating that our clustering algorithm is able to cluster a 3D data set into multiple bundles accurately. The noisy synthetic example results demonstrate the robustness of our clustering algorithm in a noisier environment.

***In Vivo* Data:** Figure 1(iii) shows the results of clustering approximately 700 trajectories from the corpus callosum into 8 bundles for two subjects. The membership probability of the trajectories for each cluster is obtained and the trajectories in Figure 1(iii) are coloured based on their maximum membership probabilities. Results showed that our clustering method automatically differentiates CC subdivision fibre bundles consistently across subjects. As a product of the proposed clustering method, regression models of each fibre bundles are obtained in the X, Y, and Z directions. Averages of these quantities are then computed over each cluster for the four subjects. The characteristics (parameters of the cubic regression equation) of each cluster are illustrated in Table 1.

Figure 2-top row show the X, Y and Z versus order U profiles for all of the tracks with mean curves for subject 1. The cluster groups are colour-coded (the same colour is used as the corresponding cluster in Figure 1 (iii)), and the mean curves for each group are highlighted in bold. Mean curves were calculated up to U=70. The mean curve results in each

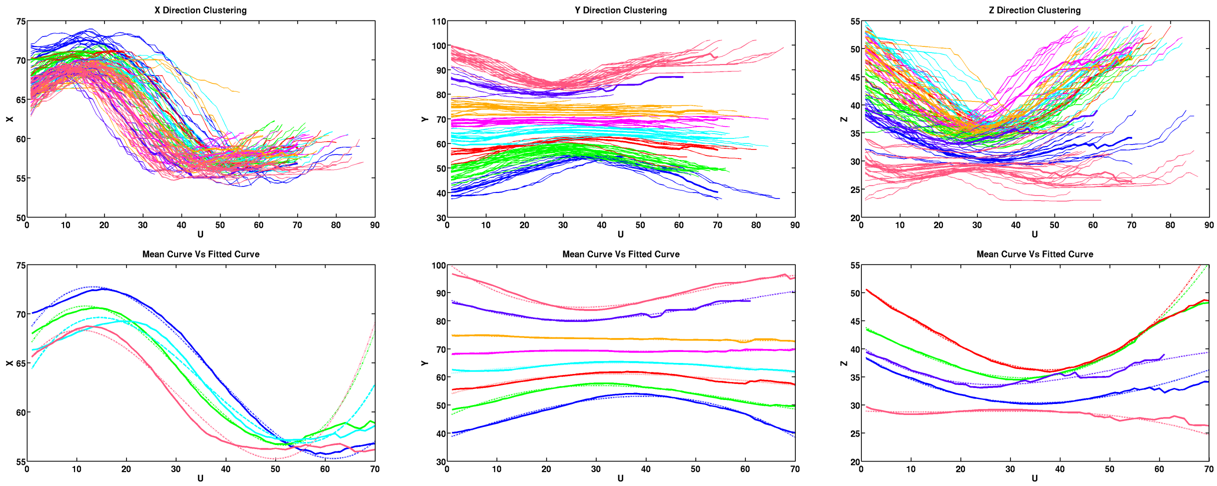


Figure 2: Top row: all the tracts, and Bottom row: the mean curves and fitted curves for the X, Y and Z directions respectively for subject 1.

direction show the fibre trajectory points, and how they each differ strongly with direction, especially the Y direction in this case. The mean curve results differ not only in shape but also in location. Figure 2-bottom row show the cubic polynomial regression models (dotted) fitted to the eight CC subdivision cluster trajectories. The results illustrate that the cubic polynomials provide the best fits among the regression models we considered.

Table 1: Cluster-wise average parameter measures for the sub-divided CC fibre bundles.

		Rostrum	Genu	Rostral body	Anterior mid body	Posterior mid body	Isthmus	Upper splenium	Lower splenium
X	β_3	3.09e-4	4.43e-4	3.46e-4	3.74e-4	4.08e-4	3.81e-4	4.08e-4	4.31e-4
	β_2	-0.0348	-0.0422	-0.0367	-0.0393	-0.0363	-0.0368	-0.0350	-0.3908
	β_1	0.7618	0.8090	0.8246	0.9103	0.6254	0.7645	0.4964	0.6804
	β_0	68.034	66.389	65.139	63.545	65.336	63.948	66.064	64.994
Y	β_3	-8.8e-5	9.3e-5	4.99e-5	-8.6e-6	3.26e-5	-1.7e-6	1.00e-4	-2.2e-4
	β_2	-0.0025	-0.0176	-0.0098	-0.0021	-0.0036	0.00053	0.0171	0.0338
	β_1	0.6171	0.8134	0.4960	0.1942	0.1275	-0.0585	-0.6694	-1.335
	β_0	38.215	45.839	53.604	61.118	67.807	74.985	87.812	100.66
Z	β_3	-3.2e-5	8.41e-5	1.35e-4	1.59e-4	-2.6e-4	4.84e-6	8.61e-5	-3.8e-5
	β_2	0.0093	0.00330	3.38e-4	5.17e-5	0.0392	0.0141	0.0138	0.00234
	β_1	-0.5317	-0.4854	-0.6009	-0.6694	-1.4661	-0.9183	-0.5589	-0.0372
	β_0	38.970	44.231	51.163	54.037	54.161	51.672	40.328	28.931

References

[1] G.Gerig, S.Gouttard, and I.Corouge. Analysis of brain white matter via fiber tract modeling. In *Proc. IEEE Int. Conf. EMBS*, 4421-4424, 2004.

[2] O'Donnell, J.Lauren, and C-F.Westin. Automatic tractography segmentation using a high dimensional white matter atlas. *IEEE Tr. Med. Im.*, 26(11):1562-1575, 2007.

[3] A.Brun, H.Knutsson, H.J.Park, M.E.Shenton, and C-F.Westin. Clustering fiber traces using normalized cuts. In *Proc. MICCAI*, 3216:368-375, 2004.

[4] M.Maddah, W.E.L.Grimson, and S.K.Warfield. A unified framework for clustering and quantitative analysis of white matter fiber tracts. *Med. Im. An.*, 12(2):191-202, 2008.

[5] S.F.Witelson. Hand and sex differences in the isthmus and genu of the human corpus callosum. *Brain*, 112:799-835, 1989.