Probabilistic Clustering and Shape Modelling of White Matter Fibre Bundles using Regression Mixtures

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 expectation-maximisation (EM) algorithm to estimate cluster membership. The result of clustering is the probabilistic assignment of fibre trajectories to each cluster and an estimate of the cluster parameters. A statistical model is calculated for each clustered fibre bundles using fitted parameters of the probabilistic clustering. We illustrate the potential of our clustering approach on synthetic data and real data.

1 Introduction

White matter (WM) fibre clustering is becoming one of the most important tasks in clinical neuroscience research and it allows to get insight about anatomical structures, perform clear visualizations and compute statistics across subjects. A number of algorithms have been developed for clustering and labelling WM fibre bundles in DTI. These algorithms can be categorised into deterministic and probabilistic approaches. Deterministic clustering algorithms assign each trajectory into only one cluster, which may lead to biased estimators of cluster parameters if the clusters overlap. Probabilistic clustering algorithms, 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.

There are a number of deterministic methods [1-3] which use different distance measures as the similarity measure among the sequence of points which parameterise each fibre tract. This family of similarity metrics deals with sets of points rather than curves and so they discard directionality information. Partial overlapping of fibres is also not taken into account as a similarity feature. Wassermann and Deriche [4] have used a publicly available anatomical atlas in conjunction with a fibre similarity metric. However, this work requires prior knowledge of the WM fibre trajectories. Maddah et al. [5] 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 in establishing correspondence between points.

While analysing fibre tracking curves geometrically is a promising notion, relatively little attention has been paid to this area, with a few exceptions. Some studies have been motivated by the problem of analysing the shapes of fibre tracts [6-7], based on a geometric framework for studying the shapes of curves in 3-D.

Corouge et al.[8] proposed a framework for quantitative tract-oriented DTI analysis that includes tensor interpolation and averaging, using nonlinear Riemannian symmetric space. In this paper, we propose a new geometrical framework to automatically cluster WM fibres into biologically meaningful neuro-tracts probabilistically. We are interested in starting with given fibre trajectories and determining whether these trajectories can be naturally clustered into groups. We investigate the model-based clustering of fibre trajectories, where each cluster is modelled as a prototype function with some variability around that prototype. A distinct feature of this model-based approach to clustering is that it produces a distinct model for each cluster. Since we are estimating smooth functions from noisy data it will be natural to use a probabilistic framework. Specifically we use mixtures of polynomial regression models as the basis of clustering. A regression model for each fibre bundle is constructed after performing probabilistic clustering.

Finite mixture models have been widely used for clustering data in a variety of application areas [5, 9-11]. Gaffney et al. employed mixtures of regression models to cluster cyclone curves [10] and hand movements [11]; although the curves were constrained to have the same length and were 2D. In contrast, we use multivariate clustering techniques to describe the three dimensional propagations of the fibre trajectories. These fibre trajectories vary in length. We use conditional mixture as it naturally allows for curves of variable length with unique measurement intervals and missing observations. The polynomial fit also takes advantage of smoothness information present in the data. Using this model, EM is performed to cluster the trajectories in a mixture model framework. The cluster membership of a particular fibre trajectory will primarily depend on how similar the trajectory shape is to each of the clusters. The algorithm is also capable of handling outliers in a principled way.

2 Probabilistic Model for White Matter Trajectories

2.1 Basic Definitions

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

$$p(v|\varphi) = \sum_{k}^{K} \alpha_k \, p_k(v|\theta_k), \tag{1}$$

in which $\varphi = \{\alpha_k; \theta_k, k = 1 \dots K\}$, α_k ($\sum_k^K \alpha_k = 1$) 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. We use the mixture model framework for fibre clustering. 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 in the k[th] 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, i.e. the cluster from which it was most likely generated.

2.2 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 $\mathrm{v_i}$. The exact form of the regression equation for $\mathrm{v_i}$ is

$$\mathrm{v_i} = \mathrm{U_i}\beta + \epsilon_\mathrm{i}, \quad \epsilon_\mathrm{i} \sim \mathrm{N}(0, \sum) \tag{2}$$

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 $\sum$.

We use the normal assumptions for ϵ_i since it is the most straightforward computationally and is the common choice for additive noise in regression models. For simplicity, we make the assumptions that $\sum = \mathrm{diag}(\sigma_x^2, \sigma_y^2, \sigma_z^2)$, so that X,Y, and Z measurement noise term are treated as conditionally independent given the model.

The conditional density for the i[th] trajectory f is a multivariate Gaussian with matrix mean $U_i\beta$ and covariance matrix $\sum$. The parameter set $\theta\ =\ \{\beta, \sum\}$.

$$\begin{aligned} p(v_i|u_i, \theta) &= f(v_i|U_i\beta, \sum) \\ &= (2\pi)^{-n_i}|\sum|^{-\frac{n_i}{2}} \exp\left\{-\frac{1}{2} tr[(v_i - U_i\beta)\sum^{-1}(v_i - U_i\beta)']\right\} \end{aligned} \tag{3}$$

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).

$$p(v_i|u_i, \varphi) = \sum_k^K \alpha_k\, f_k(v_i|U_i\beta_\mathrm{k}, \sum_\mathrm{k}) \tag{4}$$

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 $\theta_k\ =\ \{\beta_k, \sum_\mathrm{k}\}$.

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

$$L(\varphi|V, U) = p(V|U, \varphi) = \prod_i^N \sum_k^K \alpha_k\, f_k(v_i|U_i\beta_\mathrm{k}, \sum_\mathrm{k}) \tag{5}$$

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

2.3 EM Algorithm for Mixture of Regression

E-step: 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:

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

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

M-step: In the M-step, the expected cluster memberships from the E-step are used to form the weighted log-likelihood function:

$$\mathcal{L}(\varphi|V, U) = \sum_i \sum_k w_{ik} \log \alpha_k f_k(v_i|U_i\beta_k, \sum_k) \quad (7)$$

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 $W_k = diag(w'_{1k}, w'_{2k}, \ldots, w'_{nk})$ be an $N \times N$ diagonal matrix. Then, we use W_k to calculate the mixture parameters

$$\hat{\beta}_k = (U' W_k U)^{-1} U' W_k V, \quad (8)$$

$$\sum_k = \frac{(V - U\hat{\beta}_k)' W_k (V - U\hat{\beta}_k)}{\sum_1^N w_{ik}}, \quad (9)$$

and

$$\hat{\alpha}_k = \frac{1}{N} \sum_i w_{ik} \quad (10)$$

for k = 1,...,K. These update equations are equivalent to the well-known weighted least-squares solution in regression.

3 Methods

3.1 Implementation of Clustering Algorithm

Initialization: Consider a set of n three-dimensional fibre trajectories in the X, Y and Z directions. Each trajectory v_i is an n_ix3 matrix containing a sequence of n_i X, Y, and Z measurements; note that n_i may be different for each trajectory v_i. The associated n_ix1 vector of times at which the v_i measurements were observed is denoted as u_i. The number of clusters K and the order of regression p.

Algorithm 1 Regression Mixture clustering of fibre trajectories

Input: Set of 3D trajectories $V = \{v_i\}_{i=1}^N$

Output: Probabilistic assignment of trajectories to clusters.

1. Randomly initialize the membership probabilities w_{ik}
2. Calculate new estimates for parameters $\hat{\beta}_k, \hat{\sum}$ of the cluster model and mixing weights $\hat{\alpha}_k$ from Eqns (8), (9) and (10) respectively using the current w_{ik}.
3. Compute the membership probabilities w_{ik} using Eq (6) and the newly computed parameter estimates from the previous step.
4. Loop to step 2 until convergence.
5. Return the final parameter estimates (including mixing proportions) and cluster probabilities. Outliers are deleted from the set of trajectories using a threshold t.

Handling outliers: It is assumed that each trajectory is assigned a membership probability w_{ik} for each cluster k. There may be trajectories resulting from the tractography which do not resemble any of the regression equations or are not valid due to inaccuracies at the tractography stage. An outlier is identified by imposing a threshold on the membership probabilities. If the membership probability of a given trajectory in all clusters is less than the specified threshold t, that trajectory will be removed as for Maddah et al.[5].

3.2 Synthetic Data Set

The purpose of using synthetic data set is to demonstrate some of the basic features of our clustering algorithm, specifically, its ability to cluster a 3D data set into multiple bundles accurately. We used the PISTE data sets [http://cubric.psych.cf.ac.uk/commondti] which simulates different geometrical structures and several complex pathway interactions. The DTI data were simulated using a spin-echo sequence with the following parameters: number of diffusion encoding directions = 30, b-value = 1000 s/mm^2, TE = 90 ms, NEX = 4 and voxel resolution: 1x1x1 mm^3.

Here we consider three example noise free and noisy (SNR=15) data sets: a branching fibre with individual FA in each branch, two orthogonally crossing fibres with individual FA on each fibre and two straight crossing fibres. The noisy synthetic example is intended to demonstrate the robustness of our clustering algorithm in a more hostile environment–one corrupted by additive noise, with complicated fibre structures, and having varying fibre tract lengths. For the three dimensional tract reconstruction, the single-tensor and two-tensor 4th order Runge-Kutta method deterministic tractography were used for branching data and two crossing data respectively with a FA threshold of 0.15 and a curvature threshold 45^0. The generated tracts were then clustered into the subdivisions using appropriate K value. In our component regression models for the synthetic data a cubic polynomial is used. This choice is based on the visual inspection of fitted-versus-actual trajectory data.

3.3 *In Vivo* Data

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, 1.875x1.875x2.0 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 and Modelling: The corpus callosum (CC) is the largest fibre bundle that connects the two hemispheres of the brain to allow communication between the two halves of the brain. Subdividing the 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 [12] proposed a schematic for seven subdivisions of CC as shown in Fig. 1.

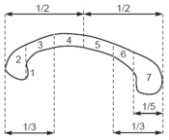

Fig 1. A schematic of Witelson corpus callosum subdivisions [12] based on the midsaggital slice: (1) rostrum (2) genu (3) rostral body (4) anterior mid-body (5) posterior mid-body (6) isthmus, and (7) splenium. We further divide the splenium into its upper and lower parts to give a finer model.

CC Trajectories were reconstructed from *in vivo* data with the fourth-order Runge-Kutta method deterministic tractography starting from the ROI specified by an expert, for all four subjects. A step size of 1mm was used. The tractography algorithm stops when it reaches a point with FA less than 0.1 or when a change in direction greater than 45^0 occurs. Fibre trajectories for the four subjects were normalized to a common template i.e. to unit voxel size (128x128x60 matrix size and voxel size 1x1x1 unit). The CC tracts were then clustered into the subdivisions using K=8.

Model Selection: It is important to make decisions about the optimal order of the fibre regression models, the most suitable type of trajectory pre-processing, and the number of clusters that best describes each fibre tract dataset for our method. We fitted regression mixture models with different orders of polynomial to randomly selected training sets of CC fibre trajectories. The experimental results were reported. 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 X position with a cubic polynomial regression model in which time 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

4.1 Synthetic Data

The Fig. 2,3 and 4 show the trajectories on original data(a), clustered trajectories(b), trajectories on noisy data (c) and clustered trajectories(d) of three selected fibre geometries. The results demonstrated the clustering algorithm to separate several fibre tracts into meaningful bundles accurately. The noise-free synthetic data results with complicated fibre tract structures demonstrate our clustering algorithm 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.

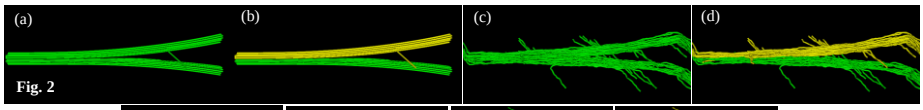

Fig. 2

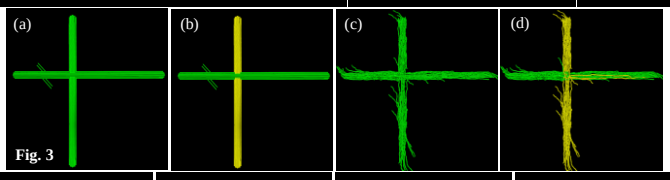

Fig. 3

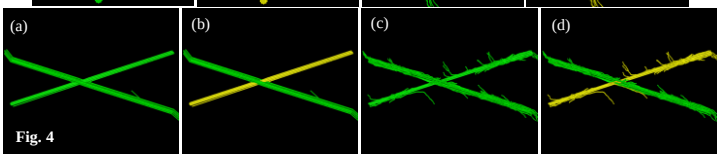

Fig. 4

4.2 *In Vivo* Data

Fig. 5 shows the results of clustering approximately 700 trajectories from the CC into 8 bundles for two subjects. The membership probability of the trajectories for each cluster is obtained and the trajectories in Fig. 5 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.

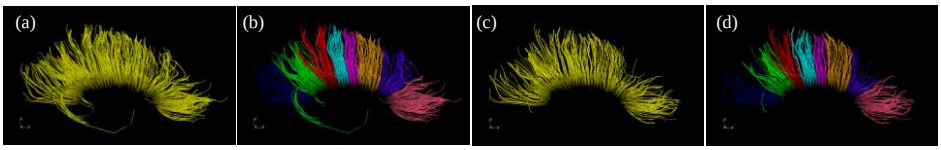

Fig. 5. Clustering of the CC from first two subjects viewed from a sagittal orientation: the original fibre tracts (a) and (c) are clustered into bundles (b) and (d) respectively.

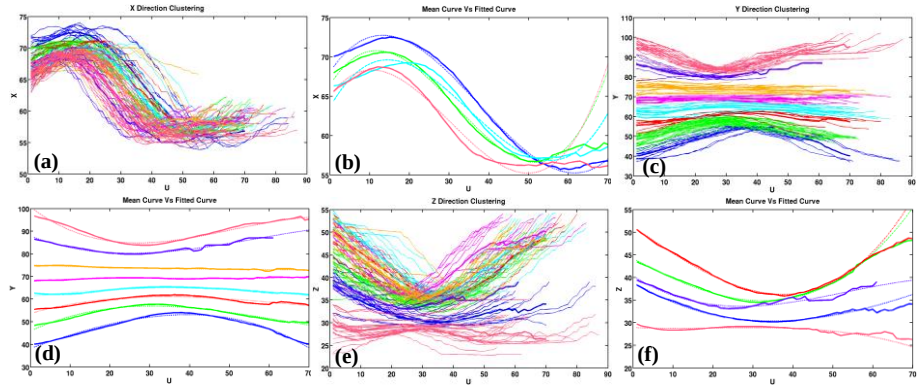

Fig. 6. (a), (c) and (e) show all the tracts, and (b), (d) and (f) are mean curves and fitted curves for the X, Y and Z directions respectively for subject 1.

Fig. 6 (a), (c) and (e) show the X, Y and Z versus order U profiles for all of the tracks with mean curves for subject 1 respectively. The cluster groups are colour-coded (the same colour is used as the corresponding cluster in Fig 5.), 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 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. Fig 4 (b), (d) and (f) 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. For each direction selected clusters are shown for clarity.

In this study, we presented new techniques for clustering 3D curves into bundles, to remove outlier curves and to develop a technique for shape description of these bundles. Curve-based regression mixture models were used to perform probabilistic clustering of fibre trajectories in three dimensional space. The number of data points is not required for clustering as the modelling can handle curves with variable lengths. The preliminary results for the synthetic data and *in vivo* data demonstrate that the new clustering process is quite efficient for bundling sets of curves into anatomically meaningful fibre tracts. Cubic polynomials were found to provide the best fits for CC clustering and modelling among the regression models considered. We have estimated cubic regression equations for each cluster fibre bundle and the equations depending on the coordinate system and image matrix, which we used.

Some of the WM trajectories are relatively small, and a successful clustering of them is heavily influenced by such factors as image quality, tractography method, and fibre tracking parameter. In the future, we will investigate how different tractography algorithms such as probabilistic tracking methods and HARDI methods affect the

WM fibre clustering procedures. The probabilistic framework allows for a variety of extensions and clustering different anatomical region which were not discussed in this paper. The number of clusters and the regression model of each anatomical region can be determined automatically. Future work directions include addressing these issues.

		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

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

References

1. Gerig, G., Gouttard, S., Corouge, I.: Analysis of brain white matter via fiber tract modeling. Proc. IEEE Int. Conf. EMBS, 4421–4424 (2004)
2. O'Donnell, Lauren, J., Westin, C-F.: Automatic tractography segmentation using a high-dimensional white matter atlas. IEEE Tr. Med. Im., 26 (11), 1562–1575 (2007)
3. Brun, A., Knutsson, H., Park, H.J., Shenton, M.E., Westin, C.-F.: Clustering fiber traces using normalized cuts. In MICCAI, 3216, 368–375 (2004)
4. Wassermann, D., Deriche, R.: Simultaneous manifold learning and clustering: Grouping WM fiber tracts using a volumetric white matter atlas. In: MICCAI Workshops (2008)
5. Maddah, M., Grimson, W.E.L., Warfield, S.K.: A unified framework for clustering and quantitative analysis of white matter fiber tracts. Med. Im. An., 12 (2), 191–202 (2008)
6. Batchelor, P.G., Calamante, F., Tournier, J.D., Atkinson, D., Hill, D.L., Connelly, A.: Quantification of the shape of fiber tracts. Magn. Reson. Med., 55, 894- 903 (2006)
7. Savadjiev, P., Campell, J.S.W., Pike, G.B, Siddiqi, K.: 3D curve inference for diffusion MRI regularization and fibre tractography. Medical Image Analysis 10, 799-813 (2006)
8. Corouge, I., Fletcher, P.T., Joshi, S., Gouttard, S., Gerig, G.: Fiber tract-oriented statistics for quantitative diffusion tensor MRI analysis. Med. Image Anal., 10, 786–798 (2006)
9. Smyth P., Ide K., Ghil M.: Multiple regimes in northern hemisphere height fields via mixture model clustering. J Atmos Sci 56(21), 3704–3723 (1999)
10. Gaffney, S. J., & Smyth, P.: Curve clustering with random effects regression mixtures. In Proceedings of the Artificial Intelligence and Statistics Workshop (2003)
11. Gaffney, S., and Smyth P.: Trajectory Clustering with Mixtures of Regression Models. In, Proc. Fifth Inter. Conf. Knowledge Discovery & Data Mining pp. 63-72 (1999)
12. Witelson S. F.: Hand and sex differences in the isthmus and genu of the human corpus callosum. Brain, vol. 112, 799–835 (1989)