NeuroImage 54 (2011) 290–302

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Tractography segmentation using a hierarchical Dirichlet processes mixture model

Xiaogang Wang [a,*], W. Eric L. Grimson [b], Carl-Fredrik Westin [c]

[a] *Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong*
[b] *Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA*
[c] *Laboratory of Mathematics in Imaging, Brigham and Women's Hospital, Department of Radiology, Harvard Medical School, Boston, MA 02215, USA*

ARTICLE INFO

Article history:
Received 2 March 2010
Revised 6 June 2010
Accepted 21 July 2010
Available online 1 August 2010

Keywords:
Tractography segmentation
Diffusion tensor imaging
Hierarchical Dirichlet process
Gibbs sampling

ABSTRACT

In this paper, we propose a new nonparametric Bayesian framework to cluster white matter fiber tracts into bundles using a hierarchical Dirichlet processes mixture (HDPM) model. The number of clusters is automatically learned driven by data with a Dirichlet process (DP) prior instead of being manually specified. After the models of bundles have been learned from training data without supervision, they can be used as priors to cluster/classify fibers of new subjects for comparison across subjects. When clustering fibers of new subjects, new clusters can be created for structures not observed in the training data. Our approach does not require computing pairwise distances between fibers and can cluster a huge set of fibers across multiple subjects. We present results on several data sets, the largest of which has more than 120,000 fibers.

Introduction

Diffusion Magnetic Resonance Imaging (dMRI) is an MRI modality that has gained tremendous popularity in recent years and is one of the first methods that made it possible to visualize and quantify the organization of white matter in the human brain in vivo. It measures local diffusivity of water within the tissue and provides information about the orientation of white matter fiber tracts. dMRI has different representations. For example, in Diffusion Tensor MRI (DT-MRI), a local 3×3 symmetric matrix is used to characterize the motion of water in all directions (Basser et al., 1994). There are more general models such as High-Angular-Resolution Diffusion-Image (HARDI) (Tuch et al., 2002), Q-ball (Tuch, 2004) and Diffusion Spectrum Imaging (DSI) (Wedeen et al., 2005) etc. Extracting connectivity information from DT-MRI, termed "tractography" (Basser et al., 2000; Conturo et al., 1999), is an especially active area of research, as it promises to model the pathways of white matter tracts in the brain, by connecting local diffusion measurements into global trace-lines. In neurological studies of white matter using tractography it is often important to identify anatomically meaningful fiber bundles. Similar fibers form clusters of points, where each cluster is identified as a "fiber bundle". Some examples are shown in Figs. 1 and 2. The fiber bundles of different subjects (examples are shown in Fig. 3) or fiber bundles of the same subject captured at different times are compared for the purposes of clinical study.

Automatically clustering fibers has drawn a lot of attention in recent years. It faces many challenges. Full brain tractography typically generates 10,000–100,000 fibers per subject. In some cases, fibers from multiple subjects need to be clustered together for comparison. Thus, the developed algorithms are expected to cluster large scale data sets. Due to data quality and other factors, tractography results may have a significant amount of errors. The clustering algorithms are required to be robust to these errors. In order to compare the fiber bundles of different subjects in group study and save computational cost, it is of interest to explore how to use the fiber bundles learned from old data sets to cluster fibers of new subjects. To obtain fiber bundles which correspond to anatomical structures, the clustering algorithms are expected to incorporate anatomical information input by experts to guide tractography segmentation.

A typical framework is to first define a pairwise similarity/distance between fibers and to input the similarity matrix to standard clustering algorithms. Various distances between fibers have been proposed. Brun et al. (2004) computed the Euclidean distances between 9-D fiber shape descriptors. Jonasson et al. (2005) measured the similarity between two fibers by counting the number of points sharing the same voxel. Maddah et al. (2005) used the *B*-spline representation to compare fibers. Gerig et al. (2004) proposed three measures related to Hausdorff distance: closest point distance, mean of closest distances and Hausdorff distance. Various clustering algorithms, such as hierarchical clustering (single-link and complete-link) (Gerig et al., 2004; Xia et al., 2005), fuzzy c-means (Maddah et al., 2008a), k-nearest neighbors (Ding et al., 2003), normalized cuts (Brun et al., 2004), spectral clustering (Brun et al., 2004; Jonasson et al., 2005; O'Donnell and Westin, 2007) and dual rooted-graphs (Tsai et al., 2007) have been

[*] Corresponding author.
E-mail addresses: xgwang@ee.cuhk.edu.hk (X. Wang), welg@csail.mit.edu (W.E.L. Grimson), westin@bwh.harvard.edu (C.-F. Westin).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.07.050

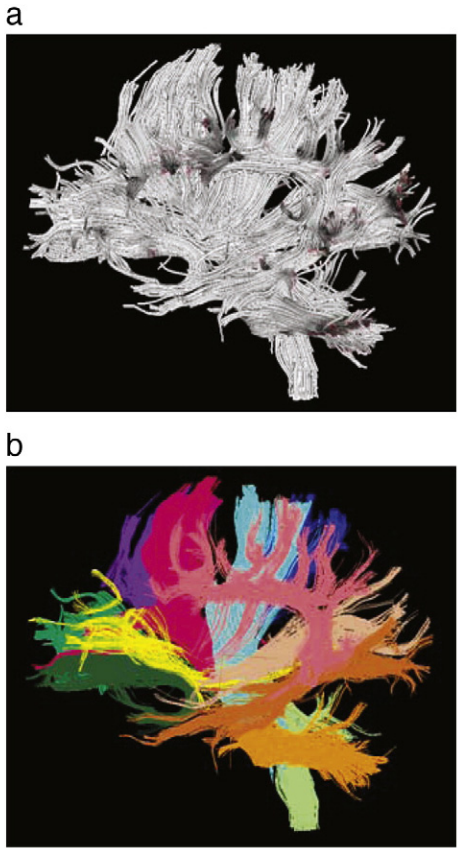

Fig. 1. Example of our tractography segmentation result. (a) Fibers generated from DT-MRI. It is hard to identify anatomical structures without clustering fibers into bundles. (b) Fiber bundles by tractography segmentation.

used. Mean of closest distances and spectral clustering are popular among possible choices (Moberts et al., 2005; O'Donnell and Westin, 2007).

These clustering algorithms required manually specifying the number of clusters or a threshold for deciding when to stop merging/splitting clusters, both of which are difficult to know especially when the data sets are complicated and noisy. Moberts et al. (2005) showed that the performance of clustering varied dramatically when different numbers of clusters were chosen. To avoid this difficulty, O'Donnell and Westin (2007) first chose a large cluster number for spectral clustering and then manually merged clusters to obtain models for white matter structures. Recently Zvitia et al. (2008) and Wassermann and Deriche (2008) used mean-shift to decide the number of clusters. Zvitia et al. (2008) required calculating the average of fibers, which was not easy in cases when fibers were not represented as feature vectors and only fiber similarities were available. Wassermann and Deriche (2008) required the white matter fiber atlas as prior knowledge. Neji et al. (2009) decided the number of clusters by adding a penalty to a larger cluster number and solving the optimization using linear programming. However, choosing the penalty term was ad hoc.

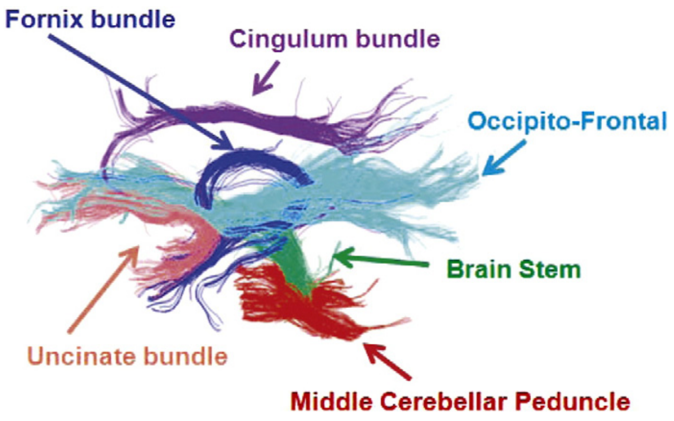

Fig. 2. Anatomical labels of some fiber bundles generated by our tractography segmentation approach.

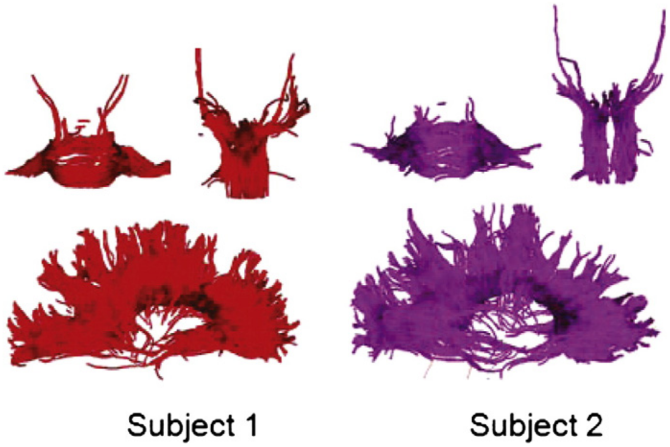

Fig. 3. Compare fiber bundles of two subjects.

Another drawback of this framework is the high space and time complexities of computing pairwise distances between fibers when the data set is large. Whole brain tractography produces between 10,000 and 100,000 fibers per subject. It is difficult to compute a 100,000 × 100,000 similarity matrix or even to store it in memory. Some clustering algorithms, such as spectral clustering, need to compute the eigenvectors of this huge similarity matrix. This problem becomes more serious when clustering fibers of multiple subjects. The current solutions are to cluster only a small portion of the whole data set after subsampling or to do some numerical approximation based on the sampled subset (O'Donnell and Westin, 2007). However, important information from the full data set may be lost after subsampling.

Besides this framework, some other approaches have been proposed in recent years (Maddah et al., 2008b; Savajiev et al., 2008; Wassermann et al., 2009; Wassermann et al., 2010). For example, Savajiev et al. (2008) proposed a fiber tract segmentation algorithm based on the geometric coherence of fiber orientations. Instead of directly grouping fiber trajectories, it clustered diffusion orientation distribution functions maxima. Maddah et al. (2008b) proposed a probabilistic approach to cluster fibers without computing pairwise distances. They used a Dirichlet distribution[1] as a prior to incorporate anatomical information. This approach is different from ours. It used a parametric model, assuming that the number of clusters is known and required manual initialization of cluster centers. Maddah et al. (2008b) required establishing point correspondence which was difficult, while our approach does not.

There are two typical ways of clustering fibers of multiple subjects for comparison purposes. One is to put fibers of different subjects into one data set and cluster them together. It has high computational cost and is not scalable. The other way is to classify the fibers of new subjects using the bundle models learned from a training set (O'Donnell and Westin, 2007). The bundle models are fixed without adaption to the new data. Among the approaches discussed above,

[1] Dirichlet distribution is used as a prior of finite mixture models. These models can only well adapt to data from particular distributions. Dirichlet process in our approach is used as a prior in infinite mixture models. These models can well adapt to a wide variety of data.

there are only a few (Maddah et al., 2005; Savajiev et al., 2008) exploring how to incorporate anatomical information from experts to guide tractography segmentation.

Nonparametric Bayesian models using Dirichlet processes (DP) as priors have been widely applied in computer vision, language processing and bioinformatics because of their capability to learn the number of clusters from data. Bayesian models involving Dirichlet process mixtures (DPM) are at the heart of the modern nonparametric Bayesian movement. Dirichlet process mixtures (DPM) models were applied to medical image analysis in recent years because of their capability to learn the number of clusters and their flexibility to adapt to a wide variety of data. Adelino and Ferreira (2006) used a DPM model for brain MRI tissue classification. In (Kim and Smyth, 2006; Thirion et al., 2007) DPM models were used to model spatial brain activation patterns in functional magnetic resonance imaging. In (Jbabdi et al., 2009), Jbabdi et al. modeled the connectivity profiles of a brain region as an infinite mixture of multivariate Gaussian distributions with a DP prior. To the best of our knowledge, our work is the first to use HDPM for tractography segmentation to automatically learn the number of clusters from data. Our approach is related to the work (Teh et al., 2006) where HDPM models were used for word-document analysis. HDPM was also used for trajectory analysis in visual surveillance (Wang et al., 2008).

In this paper, we propose a nonparametric Bayesian framework to cluster fibers into bundles. The 3D space of the brain is quantized into voxels. A bundle is modeled as a discrete distribution over voxels and orientations. This probabilistically models the spatial variation of the pathways of fibers. The models of bundles are learned from how voxels are connected by fibers instead of comparing distances between fibers. If two voxels are connected by many fibers, a bundle model will be learned with large weights on both voxels. This means that they are on the same pathway of white matter tracts. Many existing approaches have difficulty in determining the number of clusters and in clustering a very large set of fibers. Our approach automatically learns the number of clusters from data with Dirichlet processes (DP) priors (Ferguson, 1973). While the space and time complexities of existing distance-based fiber clustering approaches are at least $O(M^2)$, where M is the number of fibers, the space complexity of our approach is $O(M)$ since it does not compute and store pairwise distances between fibers.

After the models of bundles have been learnt from training data without supervision, they are used as priors to cluster/classify new fibers. Given fibers of new subjects observed, the models of bundles learned from the training set are updated and fibers of new subjects are clustered based on the updated models. Instead of fixing the number of clusters as current methods do, our approach allows the updated models to create new clusters for fiber bundles not observed in the training data. Our framework can be extended to multiscale clustering. First cluster fibers using a large size of voxels yielding bundles corresponding to structures at a large scale. Then each bundle can be further clustered using a smaller size of voxels, leading to structures at a finer scale. Multiscale clustering makes it easier for experts to identify white matter structures across different scales. Experimental evaluation on several data sets shows the effectiveness of our approach.

Method

In our method, hierarchical Bayesian models are used to cluster fibers. Hierarchical Bayesian models can provide enough parameters to fit complicated fiber bundle structures. In the meanwhile they capture the dependency among parameters through sharing priors and can better solve the overfitting problem which could be caused by a large number of parameters (Gelman et al., 2004). Moreover, hierarchical Bayesian models are flexible to be extended. For example, in Clustering new data section, our model is extended by using pre-learned bundle models as priors to guide clustering of new data. We begin by introducing a parametric hierarchical Bayesian model (Parametric model section), which is easier to understand. Using a Dirichlet process (DP) explained in the Dirichlet process section, the parametric model is extended to a nonparametric hierarchical Bayesian model, hierarchical Dirichlet process mixture (HDPM) model (Hierarchical Dirichlet processes mixture model section) and Gibbs sampling is used for inference (Inference section). In the Clustering new data section, we explain how to use the models of bundles learned from old data as a prior to cluster new data. We will use the graphical model representation to describe our probabilistic models. Readers who are not familiar with graphical models can find (Jordan, 2004) for reference.

Parametric model

Feature space

In probability theory, statistics, and machine learning, a graphical model is a graph that represents independence among random variables. The graphical model of our parametric hierarchical Bayesian model is shown in Fig. 5. There are M fibers and each fiber j has N_j points which are ordered sequentially along the fiber. Our feature space is the locations and orientations of points on fibers. $o_{ji} = (\vec{u}_{ji}, \Delta\vec{u}_{ji})$ is the observed 3D coordinate $\vec{u}_{ji} = (x_{ji}, y_{ji}, z_{ji})$ and shift $\Delta\vec{u}_{ji} = \vec{u}_{ji+1} - \vec{u}_{ji}$ of point i on fiber j. $\Delta\vec{u}_{ji}$ has the information on the orientation of point i on fiber j. The 3D space of the brain is uniformly quantized into voxels.[2] When fibers pass through a voxel, they may have different orientations. We quantize the orientations of fibers within each voxel into different directions, represented by different colors as shown in Fig. 4(a). In our experiments, shifts $\{\Delta\vec{u}_{ji}\}$ are quantized into three orientations $\Delta\vec{u}_1 = (1,0,0)^T$, $\Delta\vec{u}_2 = (0,1,0)^T$ and $\Delta\vec{u}_3 = (0,0,1)^T$. A codebook is built, in which codes (entries of the codebook) are indices of voxels and orientations. Let $\vec{u}_w$ be the centroid of the voxel and d_w be the index of the orientation vector corresponding to code w. Quantization is done in a probabilistic way,

$$p\left(o_{ji}|w\right) = p\left(\vec{u}_{ji}\,|\,\vec{u}_w\right)p\left(\Delta\vec{u}_{ji}\,|\,d_w\right), \tag{1}$$

$$p\left(\vec{u}_{ji}\,|\,\vec{u}_w\right) \propto \begin{cases} \cos^2\left(\dfrac{||\vec{u}_{ji} - \vec{u}_w||^2}{2R^2}\pi\right), & ||\vec{u}_{ji} - \vec{u}_w|| \le R \\ 0, & ||\vec{u}_{ji} - \vec{u}_w|| > R \end{cases}, \tag{2}$$

$$p\left(\Delta\vec{u}_{ji}\,|\,d_w\right) \propto \begin{cases} 1, & d_w = arg\ max_d \dfrac{|\Delta\vec{u}_{ji} \cdot \Delta\vec{u}_d|}{||\Delta\vec{u}_{ji} \cdot \Delta\vec{u}_d||} \\ 0, & \text{otherwise} \end{cases}. \tag{3}$$

R is a parameter defining a spherical neighborhood of code w. $\Delta\vec{u}_d$ is one of the three orientation vectors $\{\Delta\vec{u}_1, \Delta\vec{u}_2, \Delta\vec{u}_3\}$. As shown in Fig. 4(b), a point $\left(\vec{u}_{ji}, \Delta\vec{u}_{ji}\right)$ could be assigned to code w if it falls in the neighborhood of $\vec{u}_w$ and its shift $\Delta\vec{u}_{ji}$ is the closest to $\Delta\vec{u}_{d_w}$. The probability depends on its spatial distances to $\vec{u}_w$ and other neighboring voxels. As shown in the experimental section, the size of voxels affects the scale of fiber bundle structures found by clustering. It is chosen empirically. By choosing different sizes of voxels, multiscale clustering can be developed. By choosing the size of neighborhood R, a point could be in the neighborhoods of multiple voxels and will be assigned to one of the voxels in a probabilistic way. In practice we choose it as 1.5 times of the voxel size. Since we do not distinguish the starting and ending points of a fiber, the sign of the

[2] In this paper, "voxel" means the spatial granularity of the codebook and it is not the image voxel.

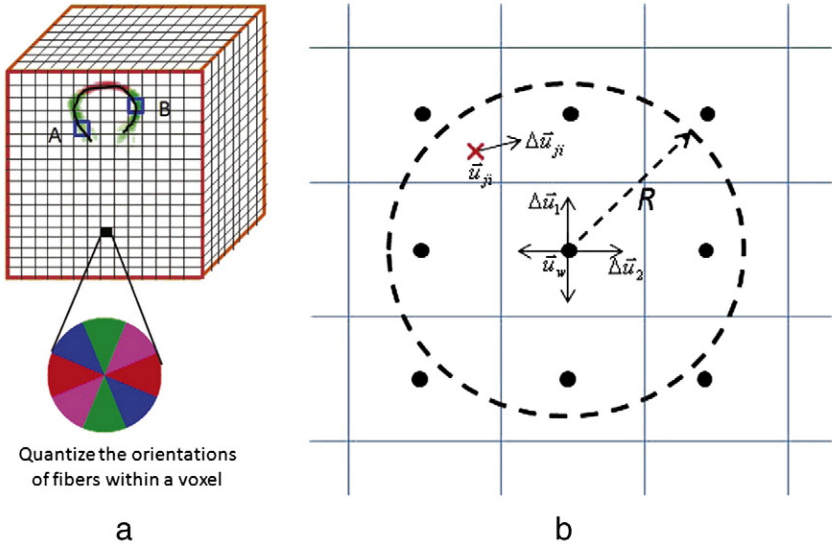

Fig. 4. Feature quantization. (a) In order to build a codebook, the 3D space of the brain is uniformly quantized into voxels. Shift orientations within voxels are also quantized (indicated by different colors). A bundle is modeled as a discrete distribution over the codebook. (b) A point $\left(\vec{u}_{ji}, \Delta\vec{u}_{ji}\right)$ on a fiber, indicated by the red cross, is quantized into a code $\left(\vec{u}_w, \Delta\vec{u}_{d_w}\right)$ according to Eqs. (2) and (3).

correlation between $\Delta\vec{u}_{ji}$ and $\Delta\vec{u}_d$ is ignored in Eq. (3). The statistical model ϕ_k of a bundle is a discrete distribution over voxels and orientations. Optionally, if the symmetry across hemispheres is considered, we can do bilateral clustering as in (O'Donnell and Westin, 2007). Assuming that the brain is aligned and $x = 0$ is the midsagittal plane, we modify observed 3D coordinates as $\vec{u}_{ji} = \left(|x_{ji}|, y_{ji}, z_{ji}\right)$ ignoring the signs of the x coordinates. Then, learnt models of bundles are symmetric to the midsagittal reflection.

The M fibers are clustered into K bundles. Each fiber bundle k is modeled as a discrete distribution ϕ_k over the codebook (i.e. quantized voxels and orientations within voxels). $\{\phi_k\}$ are learned from the co-occurrences of voxels on fibers. We assume that if two voxels are on the same bundle, they are connected by many fiber trajectories and a model ϕ_k with large distribution on both voxels is learned.

Generative model

The generative model and an example are shown in Fig. 5. H is a Dirichlet distribution over the codebook as a prior and $\{\phi_k\}$ are sampled from *Dirichlet(H)*. β is a Dirichlet distribution over bundles as a prior. Each fiber j samples a multinomial π_j from *Dirichlet*(β). Each point i on fiber j chooses one of the bundles ($c_{ji} \in \{1,\ldots,K\}$ is the bundle indicator) from a discrete distribution parameterized by π_j and samples its quantized voxel and orientation w_{ji} from the discrete distribution ($\phi_{c_{ji}}$) given by its bundle. In the given example, $c_{ji} = 1$ and

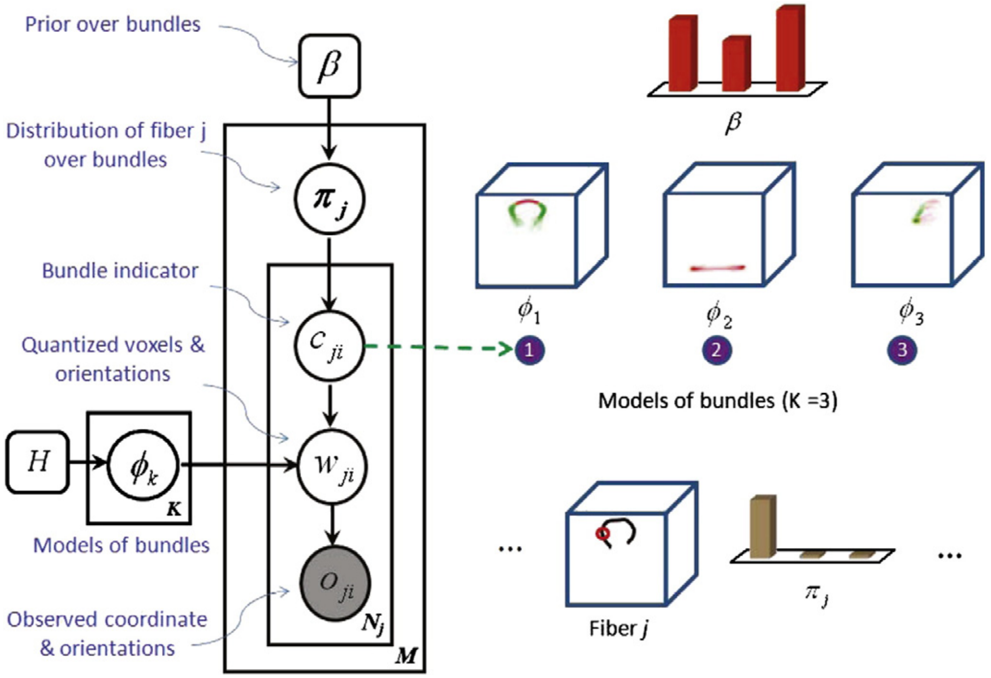

Fig. 5. The graphical model of the parametric hierarchical Bayesian model. $\{o_{ji}\}$ are the observed input. H and β are hyperparameters to be specified. $\{\phi_k\}$, $\{\pi_{ji}\}$, $\{c_{ji}\}$ and $\{w_{ji}\}$ are hidden variables to inferred by Gibbs sampling. $\{c_{ji}\}$, the labels of points on fibers, are the desired output.

w_{ji} is sampled from ϕ_1. The observed 3D coordinate and orientation o_{ji} is sampled from a kernel parameterized by w_{ji} in Eq. (1). The generative process is summarized as following.

$$\varphi_k \sim Dirichlet(H),$$
$$\pi_j \sim Dirichlet(\beta),$$
$$c_{ji} \sim Discrete\left(\pi_j\right),$$
$$w_{ji} \sim Discrete\left(\varphi_{c_{ji}}\right),$$
$$o_{ji} \sim p\left(o_{ji} | w_{ji}\right).$$

Inference

In this model, H and β are hyperparameters and are chosen as uniform distributions ($H = [h, \ldots, h]_{1 \times L}$, $\beta = [\beta_0, \ldots, \beta_0]_{1 \times K}$, where L is the size of the codebook). $\{o_{ji}\}$ are observations. $\{\phi_k\}$, $\{\pi_j\}$, $\{c_{ji}\}$ and $\{w_{ji}\}$ are hidden variables to be inferred. This model has higher data likelihood if each fiber concentrates on only a few bundles instead of uniformly distributes over all the bundles. So voxels often co-existing on the same fibers will be grouped into one bundle. In this work, collapsed Gibbs sampling is used to do inference. Collapsed Gibbs sampling integrates out some hidden variables to make the sampling converge much faster. During the sampling procedure, only $\{c_{ji}\}$ and $\{w_{ji}\}$ are sampled alternatively while hidden variables $\{\phi_k\}$ and $\{\pi_j\}$ are integrated out without being sampled. The efficiency of Gibbs sampling is significantly improved when a fewer number of variables need to be sampled. The posteriors of c_{ji} and w_{ji} are given as follows (see details of proof in Appendix A),

$$p\left(c_{ji} = k | \{c_{j'i'}\}_{j'i' \neq ji}, \left\{w_{ji}\right\}, \beta, H\right) \propto \frac{n_{jk}^{-ji} + \beta_0}{n_j^{-ji} + K\beta_0} \cdot \frac{m_{kw_{ji}}^{-ji} + h}{m_k^{-ji} + Lh}, \tag{4}$$

$$p\left(w_{ji} | o_{ji}, \{w_{j'i'}\}_{j'i' \neq ji}, \{c_{j'i'}\}_{j'i' \neq ji}, c_{ji} = k, H\right) \propto p\left(o_{ji} | w_{ji}\right) \frac{m_{kw_{ji}}^{-ji} + h}{m_k^{-ji} + Lh}. \tag{5}$$

n_j is the number of points on fiber j. n_{jk} is the number of points assigned to bundle k on fiber j. m_k is the number of points assigned to bundle k in the whole data set. m_{kw} is the number of points with voxel and orientation index w and assigned to bundle k. n_j^{-ji}, n_{jk}^{-ji}, m_k^{-ji} and m_{kw}^{-ji} are the statistics without counting point i on fiber j. Although $\{\phi_k\}$ and $\{\pi_{jk}\}$ are not explicitly sampled during the Gibbs sampling procedure, they can be estimated from any single sample,

$$\hat{\phi}_{kw} = \frac{m_{kw} + h}{m_k + Lh}, \ \hat{\pi}_{jk} = \frac{n_{jk} + \beta_0}{n_j + K\beta_0}.$$

In Eq. (4), the first term shows that point i tends to have the same bundle label as the majority of other points on the same fiber. The second term shows that the voxel and orientation index of point i needs to fit the model of the bundle chosen. From this inference procedure, we also can see that the voxels which are connected by many fibers will eventually be grouped into the same bundle.

Nonparametric model

In our parametric model, the number of bundles (K) is manually specified and it is difficult to know in advance. Using Dirichlet processes (DP) as priors, we extend the parametric Bayesian model to a nonparametric Bayesian model, which is an infinite mixture model. In this new model, the number of bundles is learned driven by data.

Dirichlet process

We first introduce DP in this section. DP (Ferguson, 1973) is used as a prior to sample probability measures. It is defined by a concentration parameter α, which is a positive scalar, and a base probability measure H. A probability measure G randomly drawn from Dirichlet process $DP(\alpha, H)$ is always a discrete distribution,

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \tag{6}$$

which can be obtained from a stick-breaking construction (Sethuraman, 1994). In Eq. (6), ϕ_k is a parameter vector sampled from H, δ_{ϕ_k} is a Dirac delta function centered at ϕ_k, and $\pi_k (\sum_{k=1}^{\infty} \pi_k = 1)$ is a non-negative scalar constructed by $\pi_k = \pi'_k \prod_{l=1}^{k-1} (1 - \pi'_l)$, $\pi'_k \sim Beta(1, \alpha)$.

G can be used as a prior for an infinite mixture model. This is called a Dirichlet process mixture (DPM) model. Let $\{w_i\}$ be a set of observed data points. Under this infinite mixture model, w_i is sampled from a density function $p(\cdot | \theta_i)$ parameterized by θ_i. θ_i (which is one of the ϕ_ks in Eq. (6)) is sampled from G. Data points sharing the same parameter vector ϕ_k are clustered together under this mixture model. Given parameter vectors $\theta_1, \ldots, \theta_N$ of N data points, the parameter vector θ_{N+1} of data point w_{N+1} can be sampled from a prior by integrating out G,

$$\theta_{N+1} | \theta_1, \ldots, \theta_N, \alpha, H \sim \sum_{k=1}^{K} \frac{n_k}{N + \alpha} \delta_{\theta_k^*} + \frac{\alpha}{N + \alpha} H. \tag{7}$$

There are K distinct parameter vectors $\{\theta_k^*\}_{k=1}^{K}$ (identifying K components[3]) among $\theta_1, \ldots, \theta_N$. n_k is the number of points with parameter vector θ_k. θ_{N+1} can be assigned as one of the existing components (w_{N+1} is assigned to one of the existing clusters) or can sample a new component from H (a new cluster is created for w_{N+1}). The posterior of θ_{N+1} is

$$p(\theta_{N+1} | w_{N+1}, \theta_1, \ldots, \theta_N, \alpha, H) \propto p(w_{N+1} | \theta_{N+1}) p(\theta_{N+1} | \theta_1, \ldots, \theta_N, \alpha, H). \tag{8}$$

It is likely for this DPM model to create a new component if existing components cannot well explain the data. There is no limit to the number of components. These properties make DP ideal for modeling data clustering problems when the number of clusters is not well-defined in advance.

Hierarchical Dirichlet processes mixture model

The extension of the parametric hierarchical Bayesian model introduced in the Parametric model section with a DP mixture model leads to our proposed hierarchical Dirichlet process mixture (HDPM) model. The graphical representation of the HDPM model is shown in Fig. 6. A prior G_0 on the whole data set is sampled from a DP, $G_0 \sim DP(\gamma, H)$, where the base measure H is a Dirichlet distribution. $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$ is a infinite mixture in which components $\{\phi_k\}_{k=1}^{\infty}$ (multinomial parameters) are models of bundles and $\beta = \{\beta_k\}_{k=1}^{\infty}$ is a prior over bundles.[4] For a fiber j, a prior G_j is sampled from a DP, $G_j = DP(\alpha, G_0)$. It was shown that in HDPM all the G_j share the same set of components $\{\phi_k\}$ as G_0. However, they have different weights π_j over $\{\phi_k\}$: $G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$, $\pi_j \sim Dirichlet(\beta)$ (Teh et al., 2006). For a point i on fiber j, a bundle indicator c_{ji} is sampled from $Discrete(\pi_j)$ and a multinomial parameters $\theta_{ji} = \phi_{c_{ji}}$ is chosen as the model of bundle c_{ji}.[5] Its index of voxel and orientation w_{ji} is sampled from the model of a bundle, $w_{ji} \sim \mathrm{Discrete}(\theta_{ji})$. Observation o_{ji} is sampled from $p(o_{ji} | w_{ji})$. Optionally, concentration

[3] When mixture models are used for clustering data, a component in a mixture model corresponds to a cluster of data. Data samples generated from the same component are grouped into the same cluster.

[4] As explained in the Dirichlet process section, $\phi_k \sim Dirichlet(H)$ and β is sampled from stick-breaking construction controlled by γ.

[5] This is equivalent to sampling θ_{ji} from G_j, $\theta_{ji} \sim G_j$.

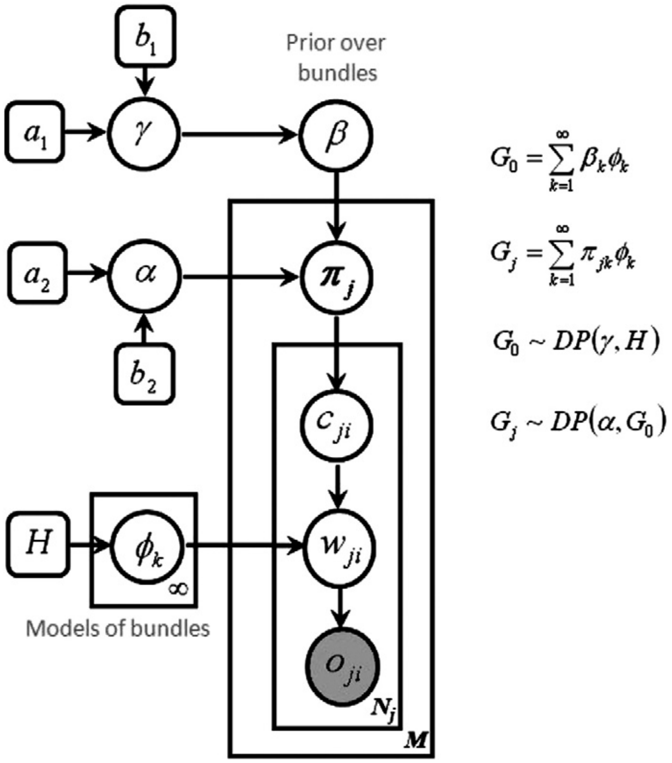

Fig. 6. Graphical model of our HDPM model. $\{o_{ji}\}$ are the observed input. a_1, b_1, a_2, b_2, and H are hyperparameters to be specified. α, β, γ, $\{\phi_k\}$, $\{\pi_j\}$, $\{c_{ji}\}$ and $\{w_{ji}\}$ are hidden variables to inferred by Gibbs sampling. $\{c_{ji}\}$ are the desired output.

parameters γ and α are sampled from gamma priors, $\gamma \sim \mathrm{Gamma}(a_1, b_1)$, $\alpha \sim \mathrm{Gamma}(a_2, b_2)$.[6] H, a_1, a_2, b_2 and b_2 are hyperparameters. The clustering performance is quite robust to the choice of their values in a large range. $\{o_{ji}\}$ are observations. The remaining are hidden variables to be inferred. A fiber j is assigned to a bundle k with maximum π_{jk}. Comparing Figs. 5 and 6, there is correspondence between parameters of the two models, except that the nonparametric model has an infinite number of bundle models $\{\phi_k\}$ and its mixture weights β and π_j have infinite components.

The size of voxels determines the scale of the structures to be learnt. Our framework can be extended to multiscale clustering. First cluster fibers using a large size of voxels so that bundles correspond to structures at a large scale. Then each bundle can be further clustered using a smaller size of voxels, showing structures at a finer scale. Multiscale clustering makes it easier for experts to identify white mater structures across different scales.

Inference

We use the collapsed Gibbs sampling method proposed in (Teh et al., 2006), which is based on Chinese restaurant franchise (Aldous, 1983), for inference. During the sampling procedure, suppose that K models of bundles (clusters) have been created and assigned to data. Then,

$$G_0 = \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u, \quad G_u \sim DP(\gamma, H). \tag{9}$$

Algorithm 1. Collapsed Gibbs sampling for our HDPM model

```
Initialization assign all the points as one cluster.
repeat
step1: sample bundle indicator $c_{ji}$ using Eq. (10).
step2: sample voxel and orientation index $w_{ji}$ using Eq. (5).
step3: sample $\beta$ using the approach proposed in (Teh et al., 2006).
step4 (optional): sample concentration parameters $\alpha$ and $\gamma$ using the approach
  proposed in (Teh et al., 2006).
until converge or exceed the maximum iteration number
```

The Gibbs sampling scheme proposed in (Teh et al., 2006) integrated out $\{\pi_{jk}\}$ and $\{\phi_k\}$ without sampling them. The posterior of c_{ji} is given by

$$p\left(c_{ji} \mid \{c_{j'i'}\}_{j'i' \neq ji}, \left\{w_{ji}\right\}, \beta, \alpha\right) \propto \begin{cases} \left(n_{jk}^{-ji} + \alpha\beta_k\right) \cdot \dfrac{n_{kw_{ji}}^{-ji} + h}{n_k^{-ji} + Lh}, & k \in \{1, \ldots, K\} \\ \alpha\beta_u \cdot \dfrac{1}{L}, & k \text{ is new} \end{cases} \tag{10}$$

This posterior allows us to create a new bundle if none of the existing bundles fits the data well. The posteriors of β, γ and α involve more details of the Chinese restaurant franchise. They can be found in (Teh et al., 2006). The posterior of w_{ji} is the same as Eq. (5). The sampling algorithm is summarized in Algorithm 1. Evaluating the convergence of Gibbs sampling is a complex issue. In practice, we stop iterations when the data log likelihood stabilizes. Let lik_t be the data log likelihood after t iterations. The stopping criterion is $\| (lik_t - lik_{(t-100)}) / lik_t \| < 0.1\%$. The maximum iteration number is set as 5000.

The space complexity of our approach is $O(M)$. The time complexity of each Gibbs sampling iteration is $O(M)$. It is difficult to provide theoretical analysis on the convergence of Gibbs sampling. In practice, we stop burn-in[7] when the data likelihood converges. From our empirical observation, the time complexity of our approach is much lower than $O(M^2)$. Recently some more efficient inference approaches, such as variational inference (Blei and Jordan, 2006), and parallel sampling (Asuncion et al., 2008), have been proposed and applied to DPM and HDPM models. In the future work, we will study how to improve the efficiency of inference using these schemes.

Clustering new data

In some applications, fiber bundles of multiple subjects need to be compared. For that purpose, we first collect a training set which includes some old subjects and learn bundle models from it in an unsupervised way. Then the pre-learned fiber bundles are used as priors to cluster fibers of each of the new subjects sequentially. Compared with alternative approaches, this approach has the following advantages. 1) If fibers of each new subject are clustered independently without using the models pre-learned from the training data, the correspondence between fiber bundles of different subjects are unknown unless it is manually specified. In our approach, the correspondence is automatically established since models of the training data are used as priors to cluster new data. 2) If both old and new subjects are merged into one data set and clustered together, although correspondence can be automatically obtained, the computational cost is high when the subject number is large. In contrast, our approach clusters new subjects one by one and saves computational cost. 3) Instead of "clustering" fibers of new subjects, alternatively one could fix the bundle models G_0 and the number of clusters learned from the training set, and use them to "classify" new fibers. However, the fixed models may not be able to fit the new data well. In our approach, pre-learned bundle models are only used as priors, new

[6] As the increase of hierarchical levels, hierarchical Bayesian models become less sensitive to hyperparameters (Gelman et al., 2004). By adding gamma priors over α and γ, α and γ do not need to be specified. Our model is more robust to the choice of a_1, a_2, b_1 and b_2 than directly tuning α and γ.

[7] Burn-in is to throw away samples generated from the early iterations of Gibbs sampling before it converges.

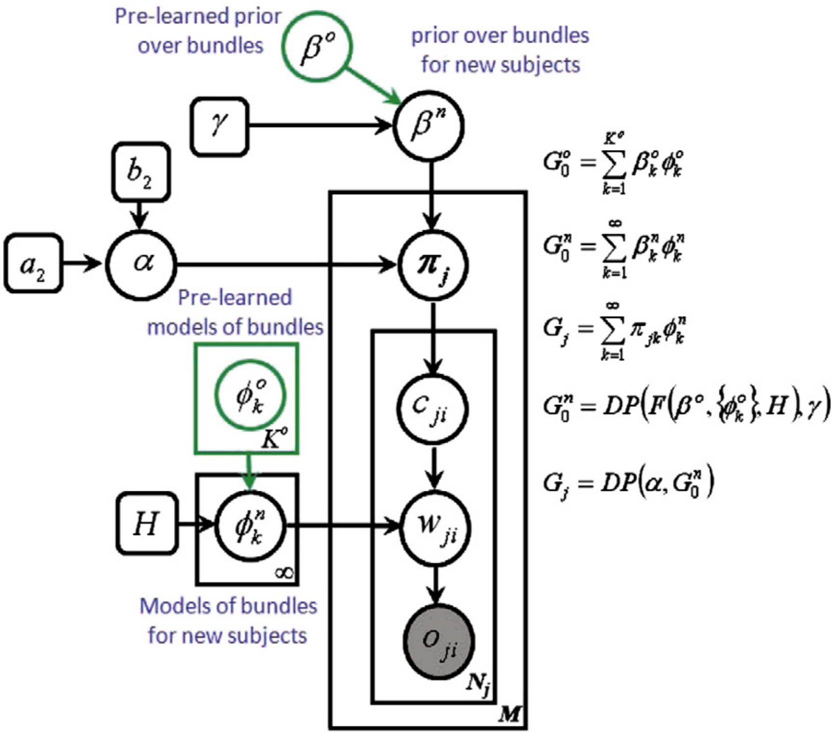
Pre-learned prior over bundles
β^o
prior over bundles for new subjects
γ
β^n
b_2
a_2
α
π_j
Pre-learned models of bundles
ϕ_k^o
K^o
c_{ji}
H
ϕ_k^n
∞
w_{ji}
Models of bundles for new subjects
o_{ji}
N_j
M
$G_0^o = \sum_{k=1}^{K^o} \beta_k^o \phi_k^o$
$G_0^n = \sum_{k=1}^{\infty} \beta_k^n \phi_k^n$
$G_j = \sum_{k=1}^{\infty} \pi_{jk} \phi_k^n$
$G_0^n = DP\left(F\left(\beta^o, \{\phi_k^o\}, H\right), \gamma\right)$
$G_j = DP\left(\alpha, G_0^n\right)$

Fig. 7. Graphical model of our HDPM model for clustering new data. $\{o_{ji}\}$ are the observed input. a_2, b_2, γ and H are hyperparameters to be specified. β° and $\{\phi_k^o\}$ are models learned from the old training data and they are fixed when clustering new data. $\{\phi_k^n\}$, β^n, $\{\pi_j\}$, $\{c_{ji}\}$ and $\{w_{ji}\}$ are hidden variables to inferred by Gibbs sampling. $\{c_{ji}\}$ are the desired output.

bundle models will be updated given the new data observed. 4) Lastly, using pre-learned models as priors can improve the convergence on the new data, since Gibbs sampling starts from a better position than random initialization.

Our HDPM model for clustering new data is shown in Fig. 7. Suppose that $G_0^o = \sum_{k=1}^{K^o} \beta_k^o \delta_{\phi_k^o} + \beta_u^o G_u^o$ in Eq. (9) has been learnt from the old data and K^o clusters are created. A prior G_0^n on the new data is to be learnt. Different from the model shown in Fig. 6, where G_0 is generated from a DP with a flat base measure H, G_0^n is generated from $DP(\gamma, F)$, where the base measure F is constructed from G_0^o and includes models learnt from the old data.

$$F = \omega \sum_{k=1}^{K^o} \hat{\beta}_k^o \delta_{\phi_k^n} + (1-\omega)H \tag{11}$$

F is composed of two parts: the models learned from the old data and a flat prior. ω is a scalar between 0 and 1. $\{\hat{\beta}_k^n\}$ are normalized weights in G_0^n,

$$\hat{\beta}_k^o = \frac{\beta_k^o}{\sum_{k'=1}^{K^o} \beta_{k'}^o}.$$

This assumes that before observing any new data, there already exist K^o models of bundles $\{\phi_k^n\}_{k=1}^{K^o}$. However, instead of letting ϕ_k^n be equal to ϕ_k^o, we sample ϕ_k^n from a Dirichlet distribution choosing ϕ_k^o as prior,

$$\phi_k^n \sim Dirichlet(\xi \cdot \phi_k^o + H),$$

where ξ_k is a positive scalar. Thus the models of bundles can adapt to the new data instead of being fixed.

The choice of γ, ω and ξ controls how much the models learned from the old data affect the clustering of the new data.[8] The two

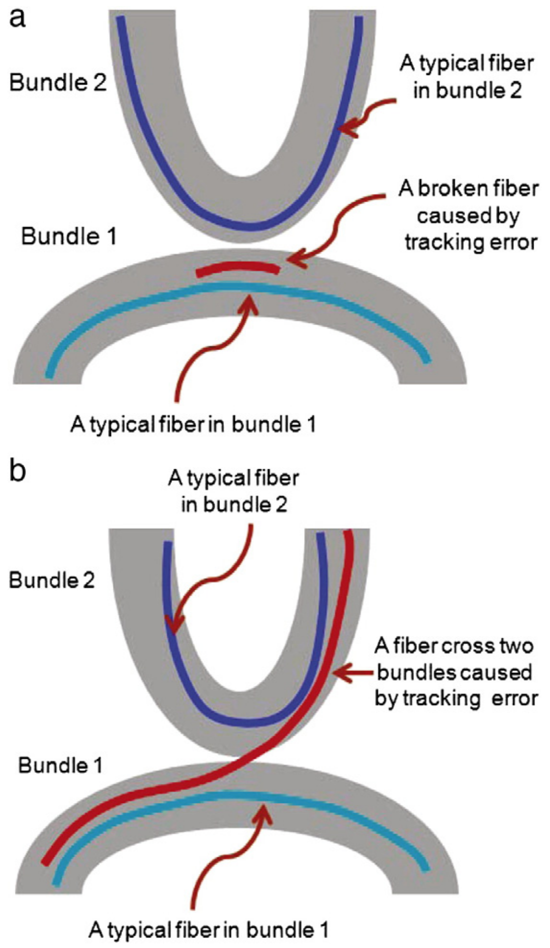
a
Bundle 2
A typical fiber in bundle 2
A broken fiber caused by tracking error
Bundle 1
A typical fiber in bundle 1
b
A typical fiber in bundle 2
Bundle 2
A fiber cross two bundles caused by tracking error
Bundle 1
A typical fiber in bundle 1

Fig. 8. Tractography errors which generate short broken fibers in (a) and fibers crossing two bundles in (b).

[8] There is a principled way to select γ, ω and ξ. See details in Chapter 4 of (Wang, 2009).

extreme cases are that the pre-learnt models have no effect on clustering the new data ($\omega=0$, $\xi=0$) and that the models learned from the new data are exactly the same as those learned from the old data ($\gamma=\infty$, $\omega=1$, $\xi=\infty$).

Suppose there are K^n models of bundles assigned to the new data. Then an explicit construction of G_0^n is given by

$$G_0^n = \sum_{k=1}^{K^o} \beta_k^n \delta_{\phi_k^n} + \sum_{k=K^o+1}^{K^n} \beta_k^n \delta_{\phi_k^n} + \beta_u^n G_{0u}^n. \tag{12}$$

Models $\{\phi_k^n\}_{k=1}^{K^o}$ have been seen in the old data. They have priors *Dirichlet*$(\xi \cdot \phi_k^o + H)$ and are updated using the new data. $\{\phi_k^n\}_{k=K^o+1}^{K^n}$ are new models not found in the old data. They are sampled from a flat prior *Dirichlet(H)*. The remaining parts are the same as described in Hierarchical Dirichlet processes mixture model section.

Discussion

Tractography results may have significant errors due to data quality and the tractography algorithms used. Two types of errors which occur frequently are shown in Fig. 8. Our tractography segmentation algorithm is more robust to these errors. In Fig. 8(a), a short broken fiber is generated because tracking terminates at a voxel of low FA value. If Euclidean distance of shape descriptors as in (Brun et al., 2004) is used, this fiber has large distance to fibers in both bundle 1 and bundle 2, because its shape is quite different from other fibers. If Hausdorff distance (O'Donnell and Westin, 2007) is used, this fiber will have small distance to fibers in both bundles, because it is close to other fibers within this local region. In our approach, each bundle has a distribution over space and orientations. Although this fiber is broken by tracking error, it better fits the model of bundle 1 than bundle 2. In Fig. 8(b), a fiber crosses two bundles because of tracking errors. Its shape is quite different from the fibers in these two bundles. The existence of this kind of fibers may lead to the two bundles merging into one cluster under some algorithms. In our approach, part of points on this fiber are assigned to bundle 1 and others are assigned to bundle 2. The models of the two bundles can be well learned without being affected by errors.

Results

Results on real data sets

We evaluate our approach on four DTI data sets. Quantitative evaluation is done on the first two data sets. The acquisition and preprocessing of the DTI data sets are the same as the *Population II* data set in (O'Donnell and Westin, 2007). DTI images were acquired using a 1.5 T MR scanner with 30 directions of diffusion weighting and $b = 700\,\mathrm{s/mm}^2$. The field of view, the size of the acquisition matrix

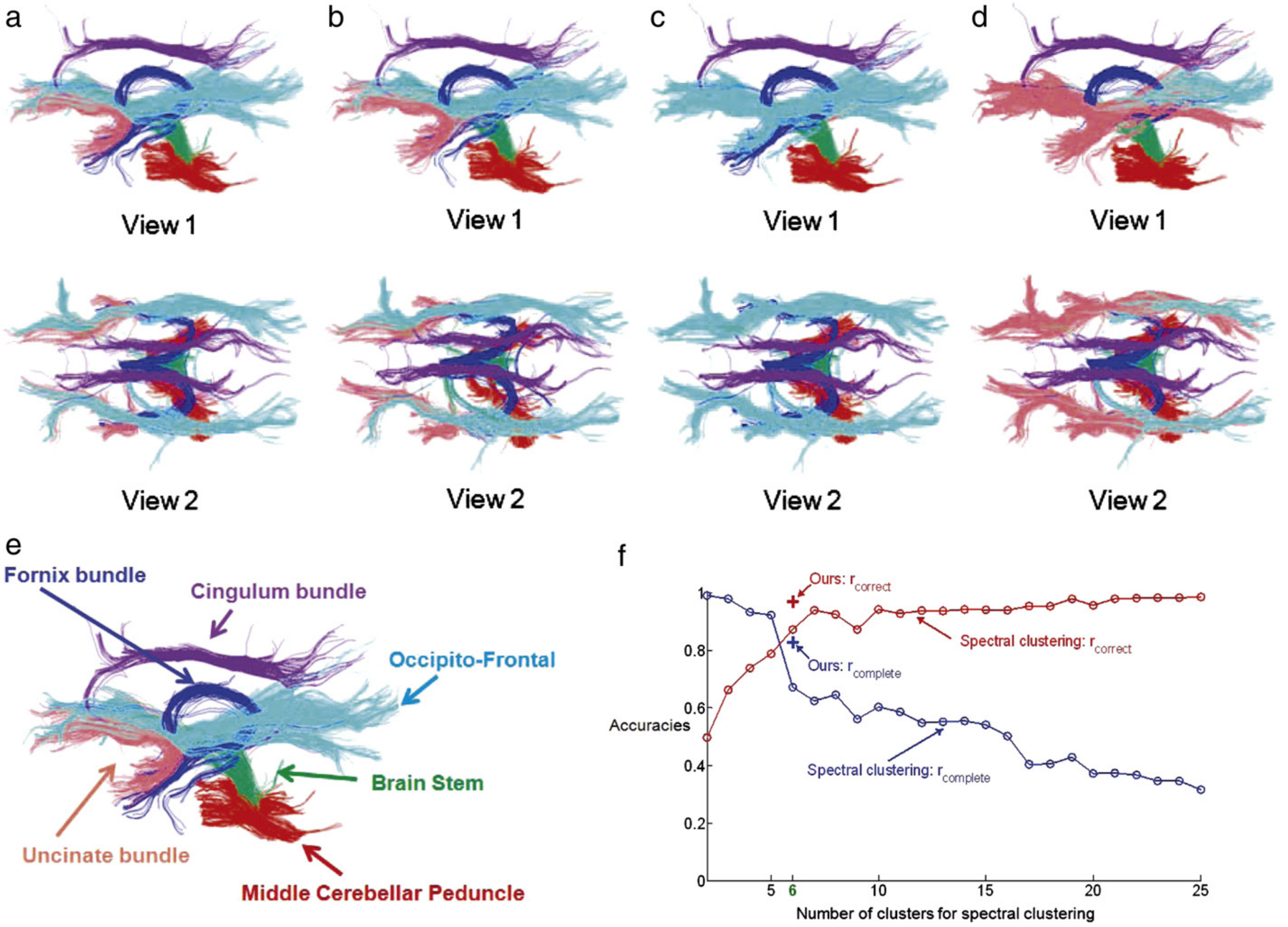

Fig. 9. Compare the results of two clustering approaches with the ground truth on a data set with 3152 fibers. Two views are plotted for each result. (a) Ground truth. (b) Our approach. (c) Spectral clustering when the number of clusters is 6. (d)Spectral clustering when number of clusters is 7. (e) Anatomical labels of the fiber bundles obtained by our approach. (f) The accuracies of completeness and correctness of spectral clustering and our approach (HDPM).

acquisition matrix, and the slice thickness were 240 mm × 240 mm/ 96 × 96/2.5 mm. Whole brain tractography was performed using Runge–Kutta order two integration, with the following parameters: seeding threshold T_{seed} of linear anisotropy measure 0.25, stopping threshold T_{stop} of linear anisotropy measure 0.15, step size 0.5 mm, and minimum length T_{length} of 25 mm. Group registration of subject FA images was performed using the congealing algorithm (Zollei et al., 2005). See more details of experimental settings in (O'Donnell and Westin, 2007). As discussed in Feature space section, 3D space of the brain is uniformly quantized into voxels. The size of each voxel is 12.5 mm × 12.5 mm × 12.5 mm. We choose the hyperparameters in Fig. 6 as $a_1 = a_2 = b_1 = b_2 = 1$, $h = 0.3$. We do bilateral clustering.

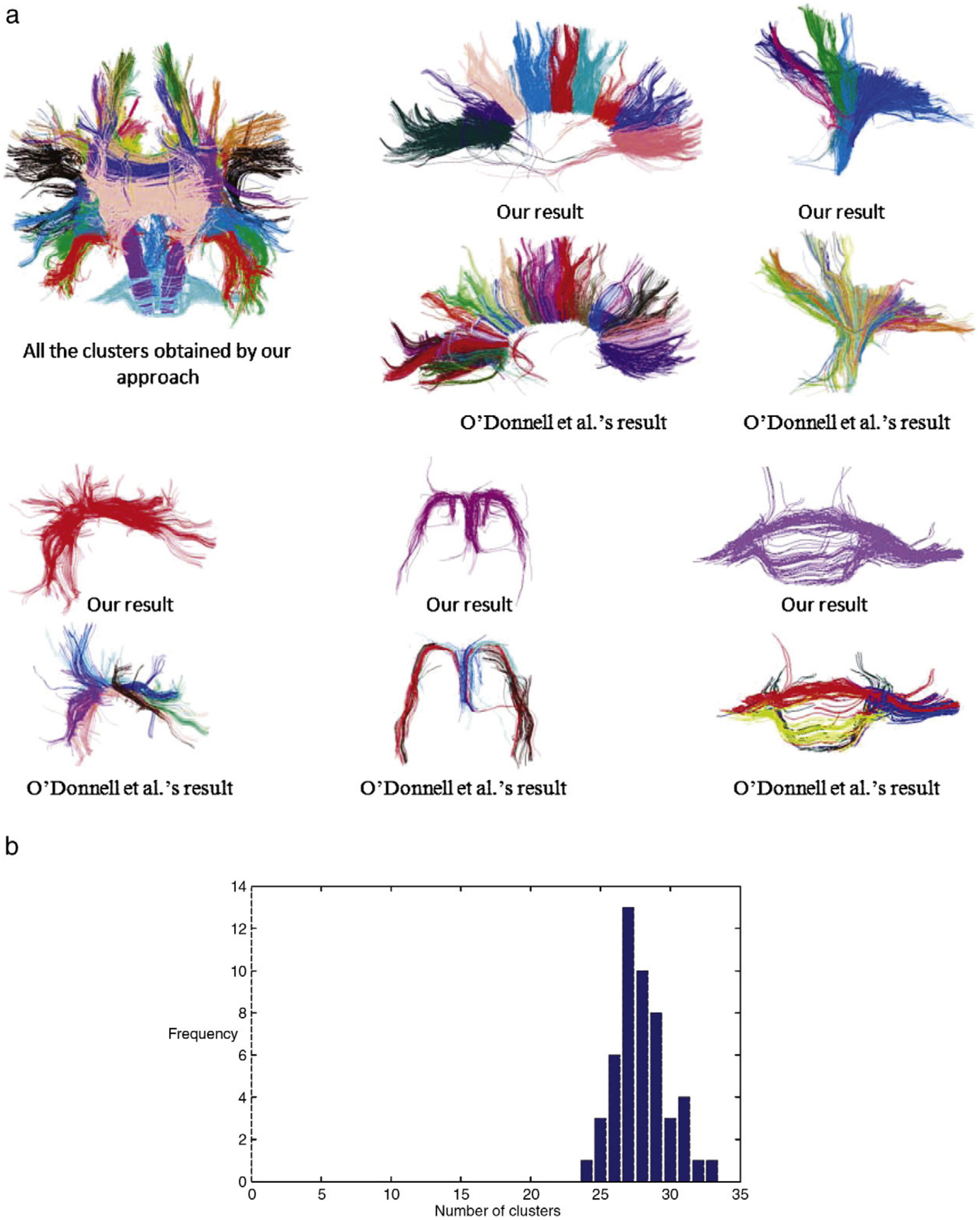
a
All the clusters obtained by our approach
Our result
Our result
O'Donnell et al.'s result
O'Donnell et al.'s result
Our result
Our result
Our result
O'Donnell et al.'s result
O'Donnell et al.'s result
O'Donnell et al.'s result
b
Frequency
0
2
4
6
8
10
12
14
0
5
10
15
20
25
30
35
Number of clusters

Fig. 10. Compared results of our approach and the approach proposed in (O'Donnell and Westin, 2007), in which experts manually merged the clusters from spectral clustering to obtain anatomical structures. (a) shows the obtained anatomical structures by merging clusters from our approach (totally 27 clusters) and those by merging clusters from spectral clustering (totally 200 clusters). Colors are used to distinguish clusters. (b) plots the frequency of the numbers of clusters learnt by our approach when running 50 trials of Gibbs sampling with random initializations.

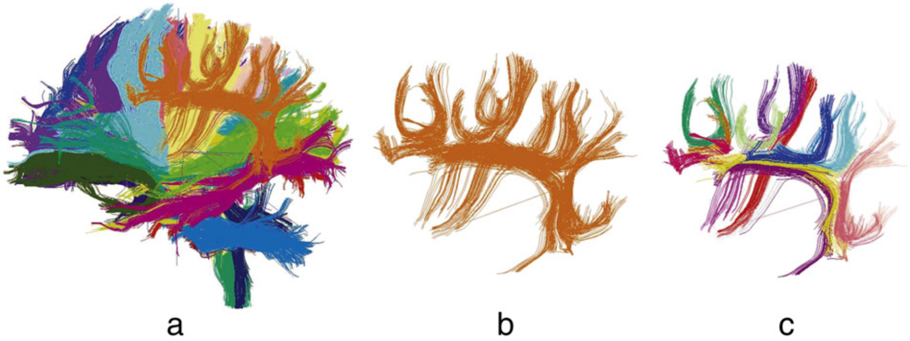

Fig. 11. An example of multiscale clustering. The spatial range of the whole brain is 240mm × 240mm × 240mm. (a): The clustering result when the space is quantized into voxels of size 12.5mm × 12.5mm × 12.5mm. The bundles correspond to structures at a large scale. (b): One bundle from (a). (c): The space is quantized into voxels of size 3.5mm × 3.5mm × 3.5mm and the bundle in (b) is further clustered into smaller bundles corresponding to structures at a finer scale.

Data set I

The first data set has 3152 fibers and is a subset of whole brain tractography results on a subject. They are manually labeled to six anatomical structures as ground truth. Figs. 9(a)–(d) plots the clustering results of our approach and a spectral clustering approach, compared with the ground truth. Colors are used to distinguish clusters. Since clusters may be permuted in different results, the meaning of colors is not consistent across different results. The spectral clustering approach uses the mean of closest distances (a variation of Hausdorff distance) as the distance measure, which was found the most effective in previous studies (Moberts et al., 2005; O'Donnell and Westin, 2007). The clustering result of our approach is close to the ground truth. Although the correct number of clusters has been set, two anatomical structures are merged in the result of the spectral clustering approach. A few outlier fibers form a small cluster. As the number of clusters increases to 7, the two anatomical structures still cannot be separated, instead, another structure splits into two clusters.

There are two important aspects, called *correctness* and *completeness*, to be considered when comparing a clustering result with the ground truth (Moberts et al., 2005). Correctness implies that fibers of different anatomical structures are not clustered together. Completeness means that fibers of the same anatomical structures are clustered together. Putting all the fibers into the same cluster results in 100% completeness and 0% correctness. Putting every fiber into a singleton cluster results in 100% correctness and 0% completeness. To measure correctness, we randomly sample 5000 pairs of fibers which are in different anatomical structures according to the ground truth and calculate the accuracy ($r_{correct}$) that they are also in different clusters according to the clustering result. To measure completeness, we randomly sample 5000 pairs of fibers which are in the same anatomical structures and calculate the accuracy ($r_{correct}$) that they

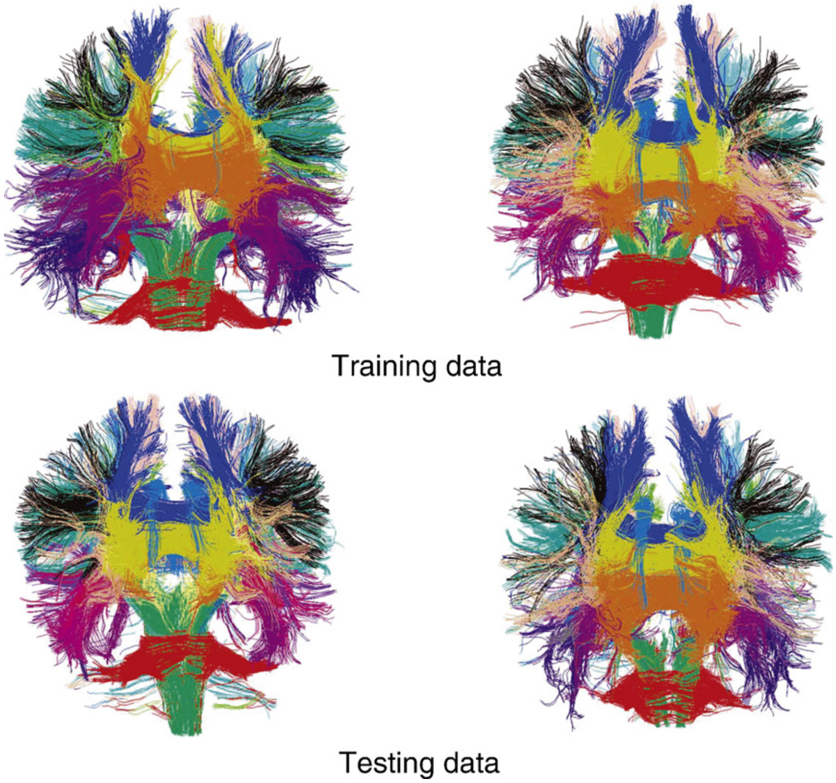

Fig. 12. Cluster fibers across multiple subjects.

are also in the same clusters. The accuracies of our approach and spectral clustering are plotted in Fig. 9(f). The horizontal axis is the number of clusters manually specified for spectral clustering. Since our approach (HDPM) automatically decides the number of clusters, which converges to six on this data set, it only has two points instead of two curves corresponding to correctness and completeness in this plot. It is observed that the correctness and completeness of spectral clustering are significantly affected by the chosen cluster number, which is hard to know in advance. As we increase the number of clusters from 2 to 25, the correctness of spectral clustering increases and its completeness decreases. When spectral clustering chooses the number of clusters as 6, which is the ground truth, both completeness and correctness of our approach are much better than those of spectral clustering. In order to achieve the same correctness as our approach, spectral clustering has to choose a cluster number larger than 19. In that case, its completeness is almost 40% lower than ours. In order to achieve a higher completeness than ours, spectral clustering has to choose fewer than 6 clusters. In that case, its correctness is almost 20% lower than ours. These observations show that our approach outperforms spectral clustering.

Data set II

We compare our approach with the approach proposed in (O'Donnell and Westin, 2007) on a larger data set with 12,420 fibers, which is also a subset of whole brain tractography results on a subject. In (O'Donnell and Westin, 2007), fibers were first grouped into a large number of clusters (200) and then experts merged these clusters to obtain anatomical structures. In this data set there are 10 anatomical structures. Our approach clusters these fibers into 27 clusters. We also manually merge them to these 10 anatomical structures, however its takes much less effort than (O'Donnell and Westin, 2007) since the number of clusters is smaller. Fig. 10 shows some of the anatomical structures obtained by the two approaches. 83.2% fibers have consistent anatomical labels according to the two results. To evaluate how our approach is sensitive to initialization, we run 50 trials of Gibbs sampling with random initializations. Fig. 10(g) plots the frequency of the numbers of clusters learnt from data.

Data set III

Fig. 11 shows an example of multiscale tractography segmentation on 30,125 fibers generated by whole brain tractography on one subject. When we choose the voxel size as 12.5 mm × 12.5 mm × 12.5 mm, the fibers obtained by full brain tractography are clustered into 35 bundles. These bundles correspond to structures at a large scale. Choosing a smaller voxel size of 3.5 mm × 3.5 mm × 3.5 mm, one of the bundles shown in Fig. 11(b) is further clustered into smaller bundles corresponding to structures at a finer scale. So far it is still not clear whether the obtained hierarchical fiber bundles well correspond to hierarchical white matter structures. It requires ground truth on white matter structures at a finer scale, which is hard to obtain at the current stage, for further evaluation. As observed in Fig. 11(b), some fibers diverging from the bundle are separated through clustering at a finer scale. They might be axons diverging from bundles and innervating the cortex. However this observation needs to be verified by further biomedical study.

Data set IV

Fig. 12 shows the results of clustering fibers across multiple subjects. The old training data has 63,751 fibers generated by whole brain tractography on two subjects. The models of bundles are learnt from all these fibers. The new data has 61,572 fibers from two new subjects. The major purposes of using the models learned from the training data as priors are to automatically establish the correspondence between bundle structures in the old and new data and to speed up the convergence. If the learned models are not used as priors, 94.2% fibers of the new data have consistent labels as using those priors.

Computational cost

Since it is difficult to provide theoretical justification on the time complexity of our approach, we provide some empirical observation on the computational cost of clustering data sets of different sizes. Running on a computer with 3 GHz CPU, it takes less than half minute to cluster 1000 fibers and around four hours to cluster 60,000 fibers. This ratio of computational costs is lower than $(60,000/1000)^2$. We empirically observe that the time complexity of our approach is lower than $O(M^2)$. The data log likelihoods of the two data sets with different number of Gibbs sampling iterations are plotted in Fig. 13. The data log likelihoods of the two data sets converge after 400 iterations and 4000 integrations respectively. With priors of pre-learned bundles models, clustering fibers of new subjects converges faster. For examples, it takes 54 min for cluster 61,572 new fibers and the data log likelihoods converge after 800 iterations.

Results on synthetic data sets

We also do experimental evaluation on fiber data synthesized by the approach proposed in (Close et al., 2009) using their default parameters. The results are shown Fig. 14. Fig. 14(b) shows the accuracies of correctness and completeness of spectral clustering and our approach when different number of bundles (from 5 to 20) are simulated. For spectral clustering, the number of clusters is manually set as ground truth. Even though our approach automatically chooses the number of clusters, it outperforms spectral clustering. The number

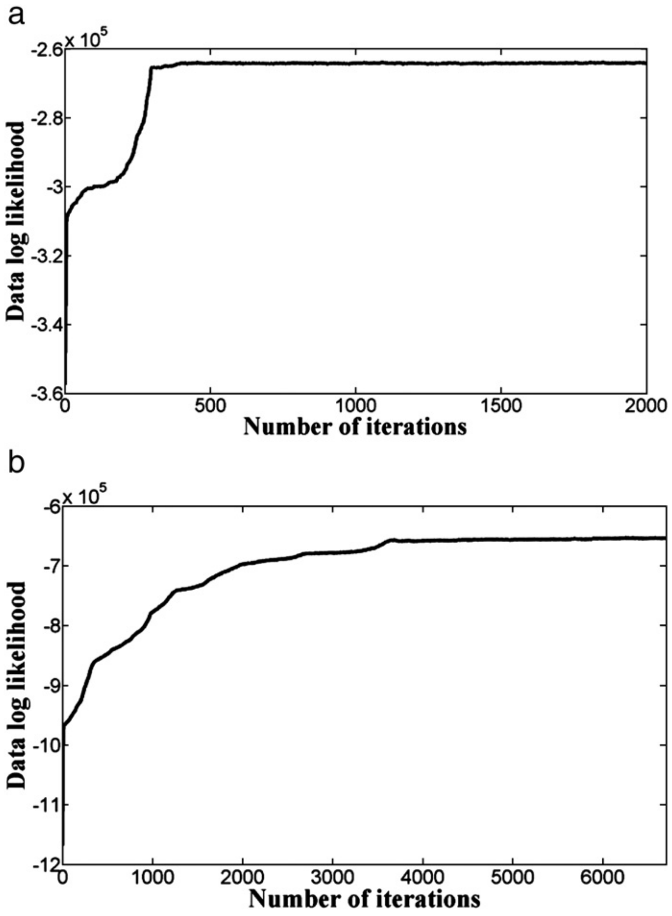

Fig. 13. Data log likelihoods of two data sets with 1000 fibers and 60,000 fibers with different number of Gibbs sampling iterations.

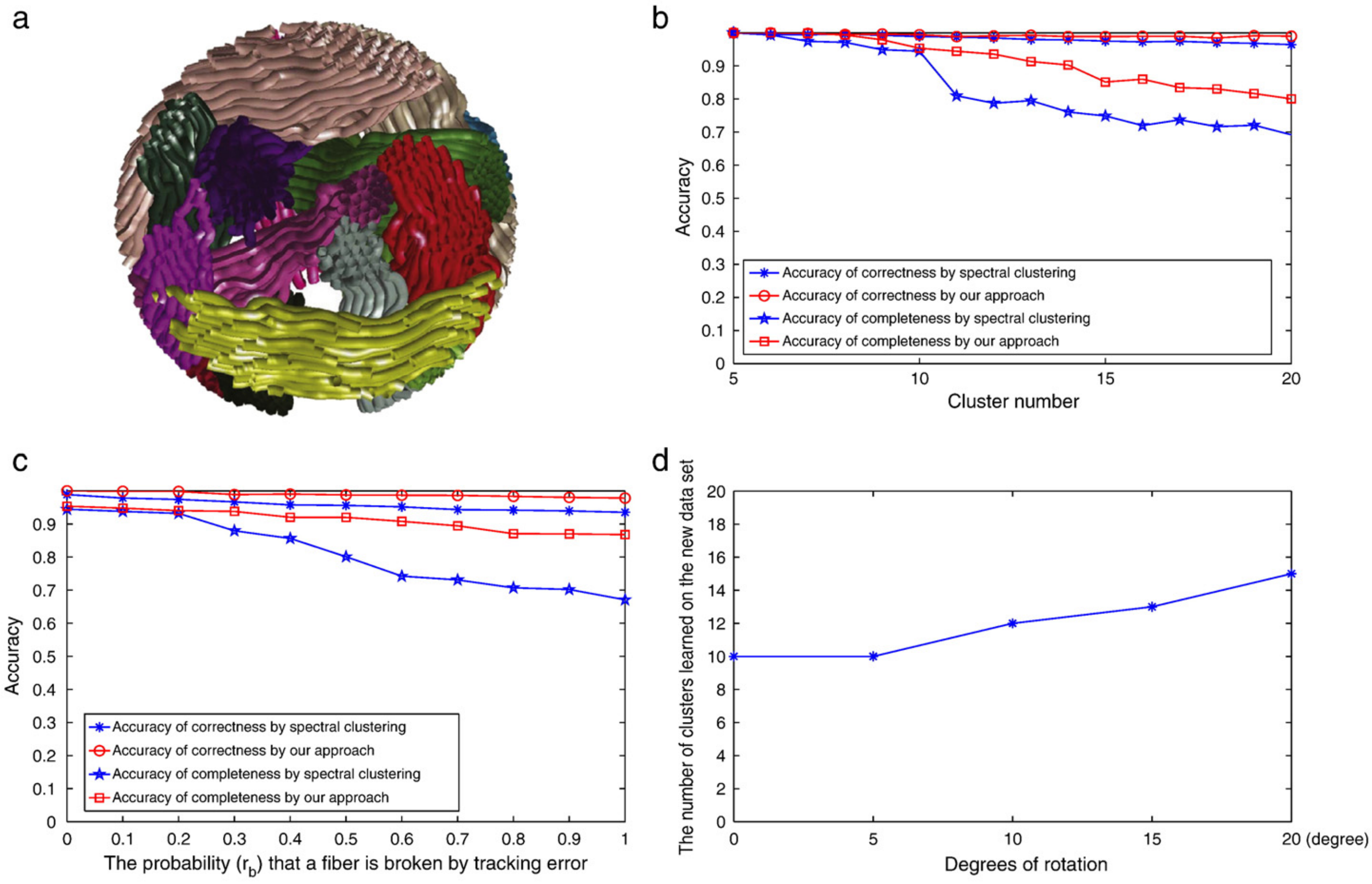

Fig. 14. Results on synthetic data. (a) Fibers synthesized by the approach proposed in (Close et al., 2009). (b) Accuracies of correctness and completeness of spectral clustering and our approach when different number of bundles are simulated. (c) Accuracies of correctness and completeness of spectral clustering and our approach when the number of bundles is ten and the simulated fibers are broken with different probabilities from 0 to 1. (d) The number of clusters learned on the new data set by our approach, when the structures of the new data set deviate from those of the training set by 0^o-20^o.

of synthesized fibers ranges from 240 to 694. As the number of bundles increases, the structures of bundles become more complicated with more overlaps, and thus the performance drops.

As discussed in Discussion section, two types of errors occur frequently in tractography results: (1) broken fibers are generated because tracking terminates at a voxel with a low FA value; (2) a fiber crosses two bundles because of tracking errors. We simulate these two types of errors by randomly splitting a fiber with probability r_b ($0 \leq r_b \leq 1$) and associating a broken fiber with a nearby broken fiber incorrectly with probability r_s ($r_s = 0.2$). As r_b increases, both the two types of errors increases. Fig. 14(c) shows that our method significantly outperforms spectral clustering with the existence of these errors.

In order to study the effect when there is mismatch between the bundle structures of the training data and the new data, we synthesize training data and new data with identical bundle structures and rotate the new data around the center by B degrees. The number of bundles is 10 in both sets. The mismatch increases with B. Fig. 14(c) shows that if the pre-learned models do not match the new data well, more clusters which do not correspond to the training data will be discovered from the new data and the cluster number on the new data becomes larger than the ground truth. In this case, the accuracies of correctness and completeness on the new data may decrease if mismatched bundle models are added as priors. For example, when $B = 20^o$, the correctness and completeness are 98.02% and 87.78% if the models learned from the training data are used as priors, while they are 99.44% and 95.35% without these priors.

Conclusion and discussion

We propose a nonparametric Bayesian framework for tractography segmentation. The number of clusters is automatically learnt from data through DP. This method has much lower space complexity than distance-based clustering methods can cluster a very large set of fibers. In the future work, we will use our model to study the groups difference between normal and diseased populations under a Bayesian framework. This approach also has some limitations. Since the clustering is based on the spatial affinity and connectivity of fiber trajectories, the biomedical principles and justification behind it are not clear. As other fully automatic clustering methods, it does not allow human intervention. In a specific application it may not be able to provide fiber bundles desired by users. In the future work, we will extend our Bayesian model to include biomedical and anatomical knowledge input by users as priors to guide tractography segmentation. When data sets are large in size, our algorithm is still not efficient enough for real time operation. The efficiency of inference can be improved using variational methods and parallel sampling.

Acknowledgments

This work was supported by NIH grants NIH R01 MH074794, NIH P41 RR13218, NIH U54 EB005149, and NIH U41 RR019703. The authors also want to acknowledge valuable discussions with Lauren O'Donnell regarding techniques for clustering fiber tracts. This research is supported by project BME-p3-10 of the Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong.

Appendix A

The posteriors of c_{ji} and w_{ji} in Eqs. (4) and (5) are given as follows. In these posteriors hidden variables $\{\pi_j\}$ and $\{\phi_k\}$ are integrated out to improve the sampling efficiency.

$$
\begin{aligned}
& p\left(c_{ji} = k | \{c_{j'i'}\}_{j'i' \neq ji}, \{w_{ji}\}, \beta, H\right) \\
& \propto p\left(c_{ji} = k, \{c_{j'i'}\}_{j'i' \neq ji}, \{w_{ji}\} | \beta, H\right) \\
& \propto \int_{\pi_j} \int_{\phi_k} p\left(c_{ji} = k, \{c_{j'i'}\}_{j'i' \neq ji}, \{w_{ji}\} | \pi_k, \phi_k, \beta, H\right) p(\pi_k, \phi_k | \beta, H) d\pi_j d_{\phi_k} \\
& \propto \int_{\pi_j} p\left(c_{ji} = k | \pi_j\right) p\left(\{c_{j'i'}\}_{j'i' \neq ji} | \pi_j\right) p\left(\pi_j | \beta\right) d\pi_j \\
& \int_{\phi_k} p\left(w_{ji} | \phi_k\right) p\left(\{w_{j'i'}\}_{c_{j'i'} = k, j'i' \neq ji} | \phi_k\right) p(\phi_k | H) d\phi_k \\
& \quad = \frac{n_{jk}^{-ji} + \beta_0}{n_j + K\beta_0} \cdot \frac{m_{kw_{ji}}^{-ji} + h}{m_k^{-ji} + Lh}
\end{aligned}
\tag{13}
$$

Since $p(\pi_j|\beta)$ (Dirichlet distribution) is a conjugate prior of $p(c_{ji}=k|\pi_j)$ and $p(\{c_{j'i'}\}_{j'i' \neq ji}|\pi_j)$ (multinomial distributions) and $p(\phi_k|H)$ (Dirichlet distribution) is a conjugate prior of $p(w_{ji}|\phi_k)$ and $p(\{w_{j'i'}\}_{c_{j'i'} = k, j'i' \neq ji}|\phi_k)$ (multinomial distributions), the two integrations in Eq. (13) has close form solutions. Similarly, we can compute the posterior of w_{ji} integrating out ϕ_k.

$$
\begin{aligned}
& p\left(w_{ji} | o_{ji}, \{w_{j'i'}\}_{j'i' \neq ji}, \{c_{j'i'}\}_{j'i' \neq ji}, c_{ji} = k, H\right) \\
& \propto p\left(o_{ji} | w_{ji}\right) p\left(w_{ji} | \{w_{j'i'}\}_{j'i' \neq ji}, \{c_{j'i'}\}_{j'i' \neq ji}, c_{ji} = k, H\right) \\
& \propto p\left(o_{ji} | w_{ji}\right) \int_{\phi_k} p\left(w_{ji} | \phi_k\right) p\left(\{w_{j'i'}\}_{c_{j'i'} = k, j'i' \neq ji} | \phi_k\right) p(\phi_k | H) d\phi_k \\
& \propto p\left(o_{ji} | w_{ji}\right) \frac{m_{c_{ji} w_{ji}}^{-ji} + h}{m_{c_{ji}}^{-ji} + Lh}.
\end{aligned}
\tag{14}
$$

References

Adelino, R., Ferreira, S., 2006. A dirichlet process mixture model for brain mri tissue classification. Med. Image Anal. 11, 169–182.

Aldous, D., 1983. Exchangeability and related topics. École d'Été de Probabilités de Saint-Flour XIII 117, 1–198.

Asuncion, A., Smyth, P., Welling, M., 2008. Asynchronous distributed learning of topic models. Proceedings of NIPS.

Basser, P.J., Mattiello, J., LeBihan, D., 1994. Mr diffusion tensor spectroscopy and imaging. Biophys. J. 66, 259–267.

Basser, P., Pajevic, S., Pierpaoli, C., Duda, J., Aldroubi, A., 2000. In vivo fiber tractography using dt-mri data. Magn. Reson. Med. 44, 625–632.

Blei, D.M., Jordan, M.I., 2006. Variational inference for Dirichlet process mixtures. Bayesain Anal. 1, 121–144.

Brun, A., Knutsson, H., Park, H.J., Shenton, M.E., Westin, C.F., 2004. Clustering fiber traces using normalized cuts. Proceedings of MICCAI.

Close, T.G., Tournier, J.D., Clamante, F., Johnston, L.A., Mareels, I., Connelly, A., 2009. A software tool to generate simulate white matter structures for the accessment of fiber-tracking algorithms. Neuroimage 47, 1288–1300.

Conturo, T.E., Lori, N.F., Cull, T.S., Akbuda, E., Snyder, A.Z., Shimony, J.S., McKinstry, R.C., Burton, H., Raichle, A.E., 1999. Tracking neuronal fiber pathways in the living human brain. Neurobiology 96, 10422–10427.

Ding, Z., Gore, J.C., Anderson, A.W., 2003. Classification amd quantification of neuronal fiber pathways using diffusion tensor mri. Magn. Reson. Med. 49, 716–721.

Ferguson, T.S., 1973. A bayesian analysis of some nonparametric problems. Ann. Stat. 1, 209–230.

Gelman, A., Carlin, H.S., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis. CRC Press.

Gerig, G., Gouttard, S., Corouge, S., 2004. Analysis of brain white matter via fiber tract modeling. Proceedings of IEEE Engineering in Medicine and Biology.

Jbabdi, S., Woolrich, M.W., Behrens, T., 2009. Multiple-subjects connectivity-based parcellation using hierarchical dirichlet process mixture models. Neuroimage 44, 373–384.

Jonasson, L., Hagmann, P., Thiran, J.P., Wedeen, V.J., 2005. Fiber tracts of high angular resolution diffusion mri are easily segmented with spectral clustering. International Society for Magnetic Resonance in Medicine.

Jordan, M.I., 2004. Graphical models. Stat. Sci. 19, 140–155.

Kim, S., Smyth, P., 2006. Hierarchical Dirichlet processes with random effects. Proceedings of NIPS.

Maddah, M., Mewes, A.U.J., Haker, S., Grimson, W.E.L., Warfield, S.K., 2005. Automated atlas-based clustering of white matter fiber tracts from dtmri. Proceedings of MICCAI.

Maddah, M., Grimson, W.E.L., Warfield, S.K., Wells, W.M., 2008a. A unified framework for clustering and quantitative analysis of white matter fiber tracts. Med. Image Anal. 12, 191–202.

Maddah, M., Zollei, L., Grimson, W.E.L., Wells III, W.M., 2008b. Modeling of anatomical information in clustering of white matter fiber trajectories using dirichlet distribution. MMBIA.

Moberts, B., Vilanova, A., Jake, J.W., 2005. Evaluation of fiber clustering methods for diffusion tensor imaging. Proceedings of IEEE Visualization.

Neji, R., Besbes, A., Komodakis, N., Deux, J.F., Maatouk, M., Rahmouni, A., Bassez, G., Fleury, G., Paragios, N., 2009. Clustering of the human skeletal muscle fibers using linear programming and angular hilbertian metrics. Porceedings of Information Processing on Medical Imaging (IPMI).

O'Donnell, L.J., Westin, C.F., 2007. Automatic tractography segmentation using a high-dimensional white matter atlas. IEEE Trans. Med. Imaging 26, 1562–1575.

Savajiev, P., Campbell, J.S.W., Pike, G.B., Siddiqi, K., 2008. Streamline flows for white matter fibre pathway segmentation in diffusion mri. Proceedings of MICCAI.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Stat. Sin. 4, 639–650.

Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical Dirichlet processes. J. Am. Stat. Assoc. 101, 1566–1581.

Thirion, B., Tucholka, A., Keller, M., Pinel, P., 2007. High level group analysis of fmri data based on Dirichlet process mixture models. Porceedings of Information Processing on Medical Imaging (IPMI).

Tsai, A., Westin, C.-F., Hero, A.O., Willsky, A.S., 2007. Fiber tract clustering on manifolds with dual rooted-graphs. IEEE International Conference on Computer Vision and Pattern Recognition.

Tuch, D.S., 2004. Q-ball imaging. Magn. Reson. Med. 52, 1358–1372.

Tuch, D.S., Reese, T.G., Wiegell, M.R., 2002. High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. Magn. Reson. Med. 48, 577–582.

Wang, X., 2009. Learning motion patterns using hierarchical bayesian models. Ph.D. thesis, Massachusetts Institute of Technology.

Wang, X., Ma, K.T., Ng, G., Grimson, W.E.L., 2008. Trajectory analysis and semantic region modeling using nonparametric bayesian models. IEEE International Conference on Computer Vision and Pattern Recognition.

Wassermann, D., Deriche, R., 2008. Simultaneous Manifold Learning and Clustering: Grouping White matter Fiber Tracts Using a Volumetric White Matter Atlas.

Wassermann, D., Bloy, L., Kanterakis, E., Verma, R., Deriche, R., 2010. Unsupervised white matter fiber clustering and tract probability map generation: Applications of a Gaussian process framework for white matter fibers. NeuroImage 51 (1), 228–241.

Wassermann, D., Bloy, L., Verma, R., Deriche, R., 2009. Bayesian framework for white matter fiber similarity measure. Proceedings of International Symposium on Biomedical Imaging.

Wedeen, V.J., Hagmann, P., Tseng, W.I., Reese, T., Weisskoff, R.M., 2005. Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. Magn. Reson. Med. 54, 1377–1386.

Xia, Y., Turken, A.U., Whitfield-Gabrieli, S.L., Gabrieli, J.D., 2005. Knowledge-based classification of neuronal fibers in entire brain. Proceedings of MICCAI.

Zollei, L., Learned-Miller, E., Grimson, W.E.L., Wells III, W.M., 2005. Efficient population registration of 3d data. Proceedings of Workshop on Computer Vision for Biomedical Image Applications.

Zvitia, O., Mayer, A., Greenspan, H., 2008. Adaptive mean-shift registration of white matter tractography. Proceedings of International Symposium on Biomedical Imaging.