White Matter Supervoxel Segmentation by Axial DP-means Clustering

Ryan P. Cabeen, David H. Laidlaw

Computer Science Department, Brown University, RI, USA

Abstract. A powerful aspect of diffusion MR imaging is the ability to reconstruct fiber orientations in brain white matter; however, the application of traditional learning algorithms is challenging due to the directional nature of the data. In this paper, we present an algorithmic approach to clustering such spatial and orientation data and apply it to brain white matter supervoxel segmentation. This approach is an extension of the DP-means algorithm to support axial data, and we present its theoretical connection to probabilistic models, including the Gaussian and Watson distributions. We evaluate our method with the analysis of synthetic data and an application to diffusion tensor atlas segmentation. We find our approach to be efficient and effective for the automatic extraction of regions of interest that respect the structure of brain white matter. The resulting supervoxel segmentation could be used to map regional anatomical changes in clinical studies or serve as a domain for more complex modeling.

Keywords: diffusion tensor imaging, atlasing, segmentation, parcellation, white matter, clustering, supervoxels, bregman divergence, directional statistics

1 Introduction and Related Work

Diffusion MR imaging enables the quantitative measurement of water molecule diffusion, which exhibits anisotropy in brain white matter due to axonal morphometry and coherence. Consequently, the orientation of fibers passing through a voxel can be estimated from the diffusion signal, allowing the local analysis of tissue or more global analysis of fiber bundles. Methods from computer vision and machine learning offer many opportunities to understand the structure captured by diffusion MRI; however, the directional nature of the data poses a challenge to traditional methods.

A successful approach for dealing with this directional data has been probabilistic models on the sphere; the two most common being the von Mises-Fisher and Watson distributions. The Watson distinguishes itself by being defined for axial variables, that is, points on the sphere with anti-podal equivalence. These models have been known in the statistics community for decades and have applications to a number of disciplines, including geophysics, chemistry, genomics, and information retrieval [16]. Recently, there has been increasing interest in exploring directional mixture models for neuroimage analysis [6, 10]. They can be

computationally challenging, however, as their normalization constants have no closed form and require either strong assumptions [11] or approximations [12].

This problem is not unique to directional data, however, and there has been much work to develop efficient and scalable alternatives to probabilistic models. Some success in this area applies to the exponential families, which constitute most distributions in use today. Banerjee et al. made a powerful finding which established a bijection between the exponential families and Bregman divergences. They also explored the asymptotic relationship between mixture models and hard clustering algorithms [2]. These hard clustering algorithms tend to be more efficient, scalable, and easy to implement, at the cost of some flexibility in data modeling. We build on these ideas to derive hard clustering for data that has both spatial and directional components. Our approach extends this idea to axial data by considering the Bregman divergergence of the Watson distribution [12]. We employ additional prior work in this area that derived the DP-means algorithm from the asymptotic limit of a Dirichlet process mixture [7, 5], providing a data-driven way to select the number of clusters. We define our segmentation from the hard clustering of voxels, so we use the terms interchangably for the rest of the paper.

Segmentation algorithms offer an opportunity to study neuroanatomy in an automated way, reducing the cost of manual delineation of anatomical structures. We consider voxelwise segmentation, as opposed to methods that cluster curves extracted by tractography. One common application of the voxelwise approach is the segmentation of the thalamic nuclei, which has been achieved by mean shift analysis, spectral clustering, level-sets, and modified k-means [17]. The work of Weigell et al. is most similar to the approach we propose, as they apply a k-means-like algorithm, modified to operate in the joint spatial-tensor domain. In contrast, we include optimization for model complexity, instead of selecting a fixed number of clusters as in k-means. We also operate on the fiber orientation, instead of the full tensor, a reasonable simplification for the segmentation of white matter, which is more anisotropic than gray matter.

We consider a segmentation of whole brain white matter in which numerous small and homogenous regions (or supervoxels) are extracted, an approach that is similar to superpixel segmentation in the computer vision literature [13]. Superpixels have been found to offer both a more natural representation of images compared to pixels and a simpler domain for more complex models [9]. In the field of biomedical imaging, this idea has recently been successfully applied to spectral label fusion [15], cellular imaging [8] and fiber orientation distribution-based segmentation [3], to which our work bears similarity.

The rest of the paper is as follows. First, we review the exponential families and their relation to Bregman divergences. We then present the models and Bregman divergences of the Gaussian and Watson distributions. From these, we derive an axial DP-means algorithm that clusters voxels in the joint spatial-axial domain. Finally, we present an evaluation with synthetic data and an application to diffusion tensor atlas white matter supervoxel segmentation.

2 Methods

2.1 Exponential Families and Bregman Divergences

In this section, we review the exponential families, its relationship to Bregman divergences, and applications of this divergence to clustering problems.

A parameteric family of distributions is considered exponential when members have a density of the following form [1]:

$$P_G(\mathbf{x}|\boldsymbol{\theta}) = P_0(\mathbf{x}) \exp(< \boldsymbol{\theta}, \mathbf{x} > -G(\boldsymbol{\theta})) \tag{1}$$

where $\boldsymbol{\theta}$ is the natural (or canonical) parameter, $\mathbf{x}$ is the sufficient statistic, $G(\boldsymbol{\theta})$ is the cumulant (or log-partition) function, and $P_0(\mathbf{x})$ is the carrier measure. The exponential families have a variety of useful properties, but here we consider their relation to Bregman divergences, a measure which can be interpreted as relative entropy. Banerjee et al. showed a bijection between the exponential families and Bregman divergences [2], and consequently, the divergence corresponding to a given member of the exponential family can be defined from the cumulant [1]:

$$\Delta_G(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = G(\hat{\boldsymbol{\theta}}) - G(\boldsymbol{\theta}) - < \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) > \tag{2}$$

This result gives a probabilistic interpretation to many distance measures. In particular, a number of hard clustering objectives may be expressed in terms of Bregman divergences associated with exponential mixture models, showing a relationship between soft and hard clustering algorithms [5]. This result can also be used to derive hard clustering algorithms for directional data, as described in the following sections.

2.2 Gaussian and Watson Distributions

We now consider two probabilistic models for our data, namely the Gaussian and Watson distributions, which represent spatial and axial data, respectively. For each distribution, we show the form of their distribution, derive their associated Bregman divergences, and discuss their relationship to hard clustering algorithms in the literature.

The isotropic Gaussian distribution is defined on $\mathbb{R}^n$ and is commonly used to represent spatial data. Its probability density $\mathcal{N}$ and associated Bregman divergence $D_{\mathcal{N}}$ are given by:

$$\mathcal{N}(\mathbf{p}|\mathbf{q}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}||\mathbf{p} - \mathbf{q}||^2\right) \tag{3}$$

$$D_{\mathcal{N}}(\hat{\mathbf{p}}, \mathbf{p}) = \frac{1}{2\sigma^2}||\hat{\mathbf{p}} - \mathbf{p}||^2 \tag{4}$$

given input positions $\mathbf{p}, \hat{\mathbf{p}} \in \mathbb{R}^n$, mean position $\mathbf{q} \in \mathbb{R}^n$, and constant variance parameter $\sigma^2 > 0$. The associated Bregman divergence $D_{\mathcal{N}}$ is the scaled Euclidean distance between two positions, an observation which gives a probabilistic interpretation to the k-means and DP-means algorithms, which are asymptotic limits of Gaussian mixture and Dirichlet process mixture models, respectively [7].

The Watson distribution is analagous to a Gaussian but defined on the hypersphere $S^{n-1} \subset \mathbb{R}^n$ with anti-podal symmetry, i.e. $\mathbf{d} \sim -\mathbf{d}$. This structure is well-suited to diffusion MR, for which a sign is not associated with the diffusion direction. Samples from this distribution are often called axial variables to distinguish them from spherical data without symmetry. Its probability density W and associated Bregman divergence D_W are given by:

$$W(\mathbf{d}|\mathbf{v}, \kappa) = \frac{\Gamma(n/2)}{(2\pi)^{n/2} M(\frac{1}{2}, \frac{n}{2}, \kappa)} \exp\left(\kappa \left(\mathbf{v}^T \mathbf{d}\right)^2\right) \tag{5}$$

$$D_W(\hat{\mathbf{d}}, \mathbf{d}) = \kappa \frac{M(\frac{1}{2}, \frac{n}{2}, \kappa)}{M'(\frac{1}{2}, \frac{n}{2}, \kappa)} \left(1 - \left(\hat{\mathbf{d}}^T \mathbf{d}\right)^2\right) \tag{6}$$

for input axial directions $\mathbf{d}, \hat{\mathbf{d}} \in S^{n-1}$, mean axial direction $\mathbf{v} \in S^{n-1}$, and Kummer's confluent hypergeometric function $M(a, b, z)$. We also assume a constant and positive concentration parameter $\kappa > 0$. The associated Bregman divergence D_W is then a scaled cosine-squared dissimilarity measure, which is equivalent to the measure used for diametrical clustering—the asymptotic limit of a mixture of Watsons [4, 12]. The Bregman divergences for the Gaussian and Watson distributions can then be used to define hard clustering algorithms, which we'll describe next.

2.3 Hard Clustering

We now present an objective function for clustering axial data and an iterative algorithm for optimizing it. In particular, this approach is a hard clustering algorithm that behaves similarly to a Dirichlet process (DP) mixture model learned with Gibbs sampling, as a result of recent work on small-variance asymptotic analysis of the exponential family and Bregman divergences by Jiang et al. [5]. We apply their work to our case of spatial and axial data, which is assumed to be modeled jointly by the Gaussian and Watson distributions.

For imaging applications, segmentation is often performed in the joint spatial-intensity space. A common application is superpixel segmentation, which provides a domain for image understanding that is both more simple and natural than the original pixel domain [9]. This is also the case for diffusion MR, where we want to segment voxels based on both proximity and fiber orientation similarity, perhaps to aid more complex anatomical modeling [3]. This motivates the development of a clustering algorithm that accounts for these two aspects, which we achieve by the linear combination of the Bregman divergences presented in the previous section. The resulting objective function E can be defined similarly to the DP-means algorithm:

$$E = \sum_{k=1}^{K} \sum_{(\mathbf{p},\mathbf{d}) \in l_k} D_{\mathcal{N}}(\mathbf{p}, \mathbf{q}_k) + D_W(\mathbf{d}, \mathbf{v}_k) + \lambda K \tag{7}$$

$$E = \sum_{k=1}^{K} \sum_{(\mathbf{p},\mathbf{d}) \in l_k} \alpha ||\mathbf{p} - \mathbf{q}_k||^2 + \beta \left(1 - \left(\mathbf{v}_k^T \mathbf{d}\right)^2\right) + \lambda K \tag{8}$$

where l_k is the set of voxels in the k-th cluster, λ is a cluster-penalty term that controls model complexity, and the parameters α and β control the relative contributions of the spatial and axial terms to the total cost. These parameters have probabilistic interpretations, where λ relates to the Dirichlet process mixture prior, and α and β relate to the Gaussian σ and Watson κ, respectively.

A procedure to minimize this objective is presented in Algorithm 1. In a modification of the DP-means algorithm [7], the assignment step measures distance by the linear combination of divergences. In the update step, the spatial cluster center $\mathbf{q}_c$ is computed by the Euclidean average. The axial cluster center $\mathbf{v}_c$ is computed by the maximum likelihood approach of Schwartzman et al. [11], where dyadic tensors are computed by the outer product of each axial variable and the mean axial direction is found from the principal eigenvector of the mean tensor, as shown by the *prineig* function.

Algorithm 1: joint spatial-axial DP-means clustering

Input:
$(\mathbf{p}_1, \mathbf{d}_1), ..., (\mathbf{p}_N, \mathbf{d}_N)$: input position/axial direction pairs,
α, β, λ: objective weighting parameters

Output:
K: number of clusters, $L_1, ..., L_N$: labels,
$(\mathbf{q}_1, \mathbf{v}_1), ..., (\mathbf{q}_K, \mathbf{v}_K)$: cluster centers

Initialize: $K \leftarrow 1,\ \mathbf{q}_1 \leftarrow \sum_i \mathbf{p}_i / N,\ \mathbf{v}_1 \leftarrow prineig\left(\sum_i \mathbf{d}_i \mathbf{d}_i^T / N\right)$
while *not converged* **do**
  Assign cluster labels:
  for *i=1* **to** N **do**
    for *j=1* **to** K **do**
      $D_{ij} \leftarrow \alpha ||\mathbf{p}_i - \mathbf{q}_j||^2 + \beta \left(1 - \left(\boldsymbol{d}_i^T \mathbf{v}_j\right)^2\right)$
    if $\min_j D_{ij} > \lambda$ **then**
      $K \leftarrow K + 1,\ \mathbf{q}_K \leftarrow \mathbf{p}_i,\ \mathbf{v}_K \leftarrow \boldsymbol{d}_i,\ L_i \leftarrow K$
    else
      $L_i \leftarrow \text{argmin}_j D_{ij}$
  Update cluster centers:
  for *j=1* **to** K **do**
    $\mathbf{q}_j \leftarrow \left(\sum_i \delta(j, L_i) \mathbf{p}_i\right) / \sum_i \delta(j, L_i)$
    $\mathbf{v}_j \leftarrow prineig\left(\left(\sum_i \delta(j, L_i) \mathbf{d}_i \mathbf{d}_i^T\right) / \sum_i \delta(j, L_i)\right)$
return

3 Experiments and Results

3.1 Synthetic Data

In our first experiment, we investigate the choice of the cluster penalty parameter λ by synthesizing axial data and testing performance of the axial DP-means algorithm across varying numbers and sizes of clusters. Here, we ignore the spatial component and only test the relationship between λ and the axial clustering. The number of clusters N ranged from 3 to 10 and were generated by sampling a Gaussian and normalizing with "size" σ ranging from 0.1 to 0.3. We evaluated ground truth agreement with the adjusted mutual information score (AMI) [14], a statistical measure of similarity between clusterings that accounts for chance groupings and takes a maximum value of one when clusterings are equivalent.

For a constant number of clusters, we found the optimal choice of λ increased with cluster size σ. For a constant cluster size σ, we found the optimal λ to be relatively stable across variable numbers of clusters N. In Fig. 1, we show examples of the clustering for the two conditions. In Fig. 2, we show plots of the relationship between the AMI and λ. We found our implementation to converge in fewer than 20 iterations on average. All cases except one found the correct number of clusters for the optimal λ. This could be due to the presence of local minima, an issue that could possibly be addressed with a randomized initialization scheme and restarts.

3.2 White Matter Segmentation

In our second experiment, we apply the axial DP-means algorithm to the supervoxel segmentation of brain white matter in a diffusion tensor atlas. We used the IXI aging brain atlas, which was constructed by deformable tensor-based registration with DTI-TK [19]. White matter was extracted by thresholding the fractional anisotropy map at a value of 0.2, which excludes white matter voxels with complex fiber configurations that are not accurately represented by tensors.

We then performed white matter segmentation of the remaining voxels with the joint spatial-axial DP-means algorithm. For each voxel, the spatial component was taken to be the voxel center and the directional component was taken to be the principal direction of the voxel's tensor. We found several structures which were spatially disconnected but given the same label, for example, the bilateral cingulum bundles, which are both close and similarly oriented. To account for this, we finally performed connected components labeling using an efficient two-pass procedure [18].

We investigated the effect of λ and β (holding α constant) by measuring the mean cluster orientation dispersion and cluster volume. We found that as λ was increased, both the volume and angular dispersion increased. As β increased, we observed decreased cluster volume and angular dispersion. These results are shown in the plots of Fig. 3. We also generated visualizations from a segmentation with $\lambda = 25$ and $\beta = 15$, which are shown in Fig. 3. From slice views, we found the segmentation to reflect known anatomical boundaries, such as the

cingulum/corpus callosum and corona radiata/superior longitudinal fasciculus. By overlaying fiber models, we see the region boundaries tended to coincide with large changes in fiber orientation. From an boundary surface rendering, we found lateral symmetry and a separation of gyral white matter from deeper white matter. Our serial implementation on a 2.3 GHz Intel i5 ran in several minutes end converged in 125 iterations.

4 Discussion and Conclusions

In this paper, we presented an efficient approach to hard clustering of spatial and axial data that is effective for segmenting brain white matter. This algorithm has a probabilistic interpretation that relates its objective to the Bregman divergences of the Gaussian and Watson distributions. Through our experiments, we first found the parameter λ to be more affected by cluster size σ than the number of clusters N. This may suggest the need to account for noise level when selecting λ and possible limitations when applied to datasets with clusters of heterogeneous size. In our second experiment, we found our approach to efficiently perform white matter atlas segmentation, producing regions that respect anatomical boundies in white matter structure. We also found considerable variablility across different hyperparameters. On one hand, this offers fine-grained control of region size, but it also suggests that a comparison with a simpler k-means type approach could be valuable. A limitation is the restriction of this approach to single fiber voxels and extensions to multiple fibers could be valuable. This could be possibly used with other methods for reconstucting fiber orientations, such as orientation distribution function (ODF) and fiber orientation distribution (FOD) approaches.

This method may also aid several aspects of clinical studies of white matter. The resulting segmentation could provide automatically defined regions-of-interest for clinical study, similar to voxel-based population studies of neurological disease. This approach may also offers a domain for more efficient inference of complex anatomical models, such as graph-based methods for measuring brain connectivity. One interesting extension would generalize the clustering objective to include variable concentration κ, which may enable more sensitive segmentation for diffusion models that account for fiber orientation dispersion in each voxel. In conclusion, we find this approach to be an efficient and valuable tool for segmenting white matter with a desirable probabilistic interpretation and a number of applications to brain connectivity mapping.

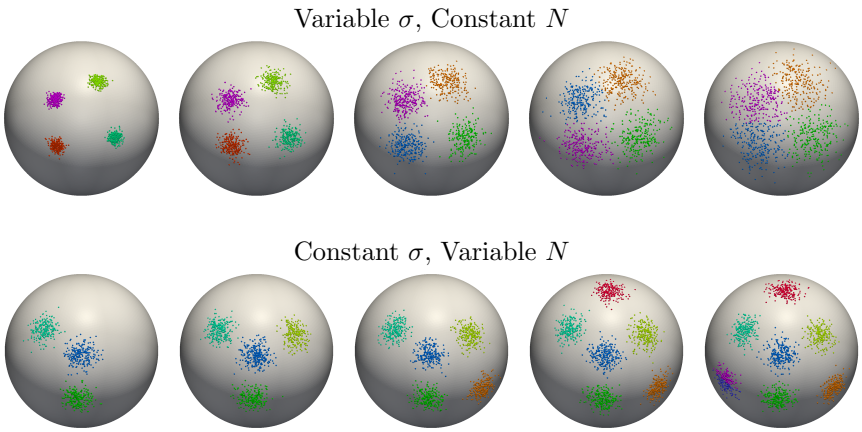

Fig. 1: First experiment: visualizations of the synthetic axial data and the optimal clusterings computed from the proposed method. The top shows results for variable cluster size $\sigma = \{0.05, 0.0875, 0.125, 0.1625, 0.2\}$, and constant number of clusters $N = 4$. The bottom shows results for constant cluster size $\sigma = 0.10$, and variable number of clusters $N = \{3, 4, 5, 6, 7\}$. The bottom right shows a single mislabeled cluster, possibly caused by finding a local minima in the optimization.

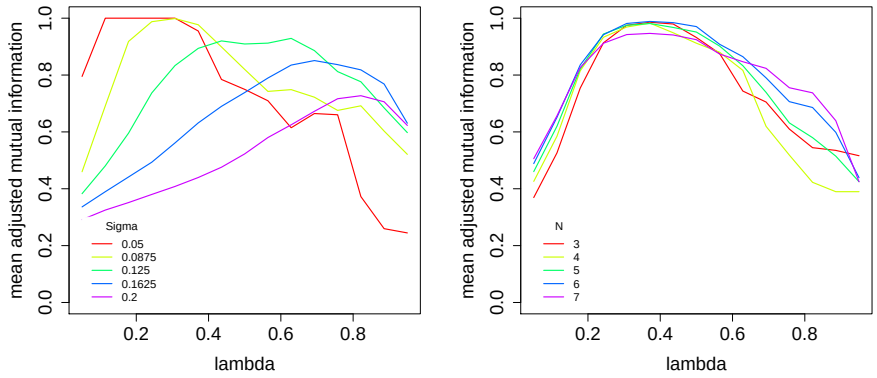

Fig. 2: First experiment: clustering performance as a function of cluster penalty parameter $\lambda \in [0, 1]$ given ground truth generated with cluster size σ, and number of clusters N, which are visualized in Fig. 1. We measured the adjusted mutual information (AMI), a statistical measure takes a maximal value when clusterings are equivalent. Shown are plots of the AMI vs. λ for two conditions. The first tested with variable $\sigma = \{0.05, 0.0875, 0.125, 0.1625, 0.2\}$ and constant $N = 4$. The second tested with constant $\sigma = 0.10$ and variable $N = \{3, 4, 5, 6, 7\}$. These results indicate that the optimal λ depends more on σ than N, suggesting that performance may depend on the noise level and may dimishish when differing clusters sizes are present. Also note that the maximum AMI decreases with increasing σ, which may be due to cluster overlap or over-sensitivity in the AMI measure.

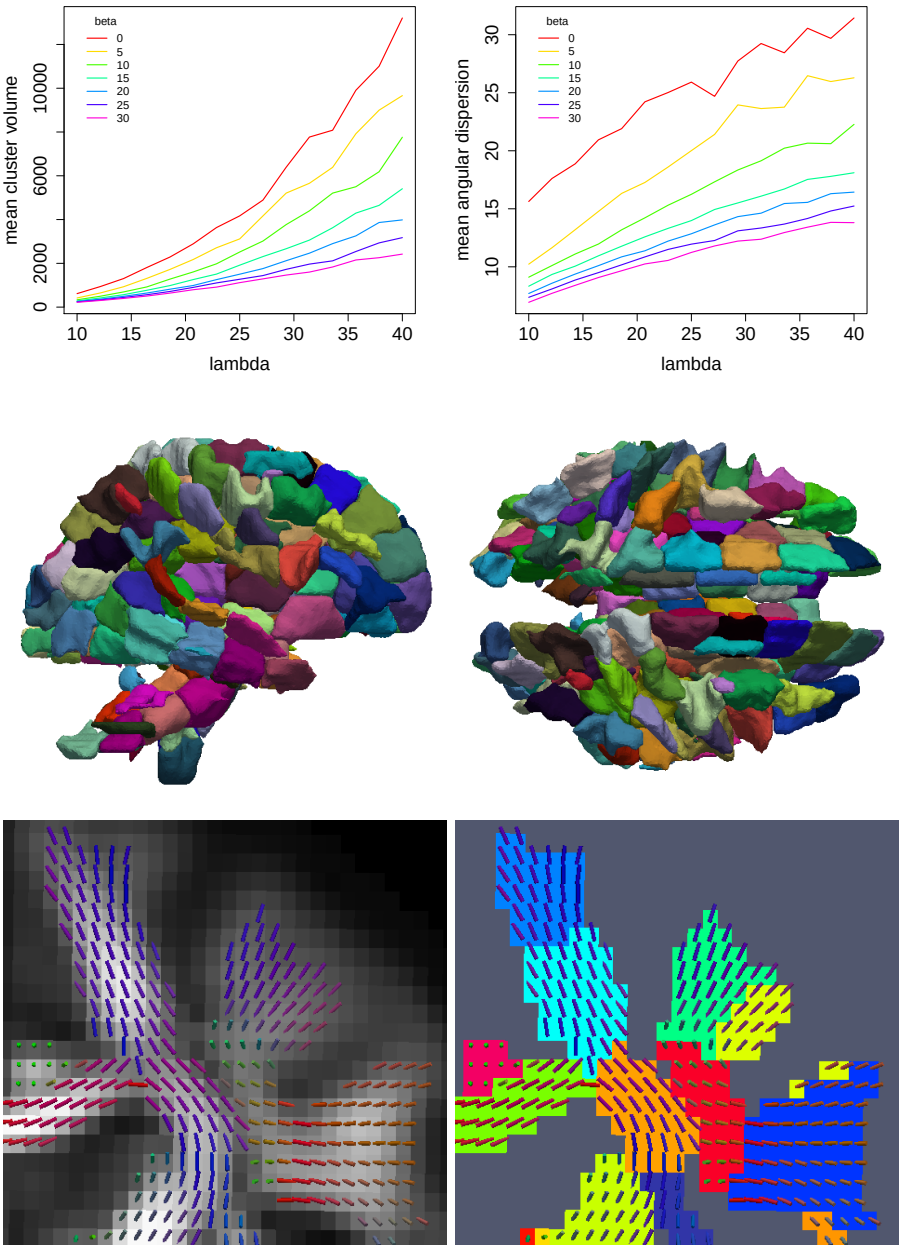

Fig. 3: Second experiment: white matter supervoxel segmentation by axial DP-means clustering. The top row shows mean cluster volume (mm^3) and mean cluster angular dispersion (degrees) as a function of cluster penalty $\lambda \in [10, 40]$ and axial weighting $\beta \in [0, 30]$. We found increasing λ also increased the volume and dispersion, while increasing β reduced the volume and dispersion. The middle and bottom rows show an example result given $\lambda = 25$ and $\beta = 15$. The middle shows boundary surfaces of the regions, which illustrates the symmetry and separation of gyral and deep white matter. The bottom shows a partial coronal slice with the fractional anisotropy (left) and computed segmentation (right), whish shows region boundaries that match known anatomical interfaces, such as the corpus callosum/cingulum bundle and corona radiata/superior longitudinal fasciculus.

References

1. Azoury, K.S., Warmuth, M.K.: Relative loss bounds for on-line density estimation with the exponential family of distributions. Machine Learning 43(3) (2001)
2. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. J. Mach. Learn. Res. 6, 1705–1749 (Dec 2005)
3. Bloy, L., Ingalhalikar, M., Eavani, H., Schultz, R.T., Roberts, T.P., Verma, R.: White matter atlas generation using HARDI based automated parcellation. NeuroImage 59(4), 4055 – 4063 (2012)
4. Dhillon, I.S., Marcotte, E.M., Roshan, U.: Diametrical clustering for identifying anti-correlated gene clusters. Bioinformatics 19(13), 1612–1619 (2003)
5. Jiang, K., Kulis, B., Jordan, M.: Small-variance asymptotics for exponential family Dirichlet process mixture models. In: NIPS 2012 (2012)
6. Kaden, E., Kruggel, F.: Nonparametric Bayesian inference of the fiber orientation distribution from diffusion-weighted MR images. Med. Image Anal. 16(4), 876 – 888 (2012)
7. Kulis, B., Jordan, M.I.: Revisiting k-means: New algorithms via Bayesian nonparametrics. In: ICML-12. pp. 513–520 (2012)
8. Lucchi, A., Smith, K., Achanta, R., Lepetit, V., Fua, P.: A fully automated approach to segmentation of irregularly shaped cellular structures in em images. In: Jiang, T., Navab, N., Pluim, J., Viergever, M. (eds.) MICCAI 2010, LNCS, vol. 6362, pp. 463–471. Springer Berlin Heidelberg (2010)
9. Mori, G.: Guiding model search using segmentation. In: ICCV 2005. vol. 2, pp. 1417–1423 Vol. 2 (2005)
10. Rathi, Y., Michailovich, O., Shenton, M.E., Bouix, S.: Directional functions for orientation distribution estimation. Med. Image Anal. 13(3), 432 – 444 (2009)
11. Schwartzman, A., Dougherty, R.F., Taylor, J.E.: Cross-subject comparison of principal diffusion direction maps. Magnet. Reson. in Med. 53(6), 1423–1431 (2005)
12. Sra, S., Karp, D.: The multivariate Watson distribution: Maximum-likelihood estimation and other aspects. J. of Multivariate Analysis 114(0), 256 – 269 (2013)
13. Veksler, O., Boykov, Y., Mehrani, P.: Superpixels and supervoxels in an energy optimization framework. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, LNCS, vol. 6315, pp. 211–224. Springer Berlin Heidelberg (2010)
14. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res. 11, 2837–2854 (Dec 2010)
15. Wachinger, C., Golland, P.: Spectral label fusion. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, LNCS, vol. 7512, pp. 410–417. Springer Berlin Heidelberg (2012)
16. Watson, G.S.: Statistics on spheres, vol. 6. Wiley New York (1983)
17. Wiegell, M.R., Tuch, D.S., Larsson, H.B., Wedeen, V.J.: Automatic segmentation of thalamic nuclei from diffusion tensor magnetic resonance imaging. NeuroImage 19(2), 391 – 401 (2003)
18. Wu, K., Otoo, E., Suzuki, K.: Optimizing two-pass connected-component labeling algorithms. Pattern Analysis and Applications 12(2), 117–135 (2009)
19. Zhang, H., Yushkevich, P., Rueckert, D., Gee, J.: A computational white matter atlas for aging with surface-based representation of fasciculi. In: Fischer, B., Dawant, B., Lorenz, C. (eds.) Biomedical Image Registration, LNCS, vol. 6204, pp. 83–90. Springer Berlin Heidelberg (2010)