The Approximation of the Dissimilarity Projection

Emanuele Olivetti[1], Thien Bao Nguyen[1]
Eleftherios Garyfallidis[2]

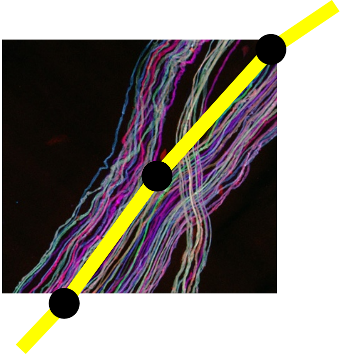
[1]NeuroInformatics Laboratory (NILab)
Bruno Kessler Foundation, Trento (FBK), Italy
Center for Mind and Brain Sciences (CIMeC), University of Trento, Italy
http://nilab.fbk.eu
olivetti@fbk.eu
[2]MRC Cognition and Brain Sciences Unit, University of Cambridge, UK

2nd International Workshop on Pattern Recognition in Neuroimaging, July 2-4 2012, UCL, London, UK

Streamlines

Basics

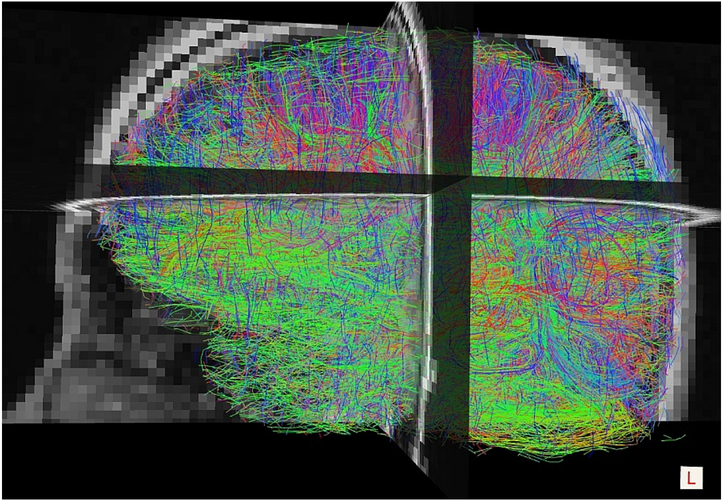
- dMRI techniques allow the reconstruction of pathways in living subjects. Res. $\approx 2mm$.
- Tractography algorithms reconstruct streamlines/*fibers*.
- A streamline is a polyline representing thousands of axons.



Notation

- Streamline: a polyline $X = \{\mathbf{x_1}, \dots, \mathbf{x}_{n_X}\}$, where $\mathbf{x} \in \mathbb{R}^3$.
- Tractography: $S = \{X_1, \dots, X_N\}$. Usually $|S| \simeq 3 \times 10^5$.

Tractography: $\approx 3 \times 10^5$ streaml. Here: 5%.

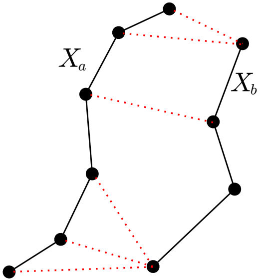


Tractography data and Pat.Rec./Mach.Learn.

- **Pros:** distance [Zhang et al., 2008] between streamlines:
  $d(X_a, X_b) = \frac{1}{2}(\delta(X_a, X_b) + \delta(X_b, X_a))$

  $$\delta(X_a, X_b) = \frac{1}{|X_a|} \sum_{\mathbf{x}_i \in X_a} \min_{\mathbf{y} \in X_b} ||\mathbf{x}_i - \mathbf{y}||_2.$$

- **Cons:** streamlines have different lengths / number of points.



How to do Classif./Cluster. on Tractography Data? [Olivetti and Avesani, 2011]

The *Dissimilarity Representation* [Pekalska et al., 2002]: a Euclidean embedding from the Pat.Rec/ML literature.

Today's questions

- **How accurate is the Dissimilarity Projection?**
- **How to efficiently select the prototypes?**

Outline

1. The Dissimilarity Projection/Representation.
2. Prototype selection algortihms:
   - Farthest First Traversal
   - Subset Farthest First
3. A measure of the degree of approximation.
4. Experimental results.
5. Conclusions.

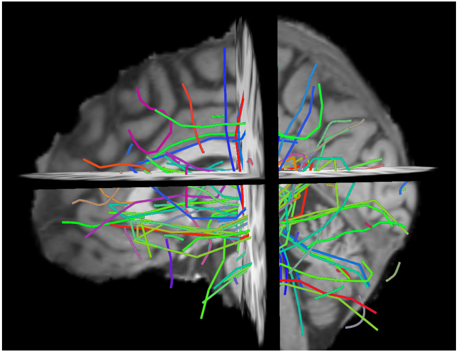
Euclidean Embedding: *The Dissimilarity Projection*

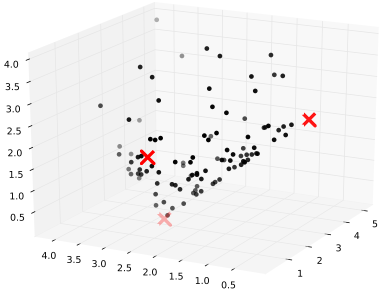
1. Select a set of p streamlines (*prototypes*)

$$\Pi = \{\tilde{X}_1, \ldots, \tilde{X}_p\}$$

2. Each new streamlines is represented as the vector of distances to the prototypes

$$\phi_\Pi^d(X) = [d(X, \tilde{X}_1), \ldots, d(X, \tilde{X}_p)]$$





How to Select Prototypes? *Farthest First Traversal*

"The optimal solution to the k-center problem is NP-hard."
*"The **Farthest First Traversal** (FFT) algorithm is optimal among non-NP-hard solutions."* [Hochbaum and Shmoys, 1985]

FFT algorithm

1. $\tilde{X}_1$: select one streamline at random.
2. $\tilde{X}_{i+1}$ is the farthest streamline from all previously selected.

In [Pekalska et al., 2006] FFT is shown to be very accurate for classification problems.
Scalability Issue: $O(p|S|)$ evaluations of $d(X_a, X_b)$.

- Example: if $p = 30$ and $|S| = 3 \times 10^5$, then $\approx 10^7$ evaluations.

How to Select Prototypes? *Subset Farthest First*

In [Turnbull and Elkan, 2005] it is proved that:

Subset Farthest First

1. Sample $m = \lceil cp \log p \rceil$ streamlines from S at random.
2. Select the prototypes from this sample with FFT.

Lemma: *"under the hypothesis of p clusters in S, the probability of not having a representative of some clusters in the sample is* $< pe^{-m/p}$*"*.

- Example: if $p = 30$, $c = 3$ then $m = 307$, prob< 0.001

Complexity: $O(cp^2 \log p)$ evaluations of $d(X_a, X_b)$.

Independent of $|S|$!!

- Example: if $p = 30$, prob< 0.001, then $\approx 10^4$ evaluations.

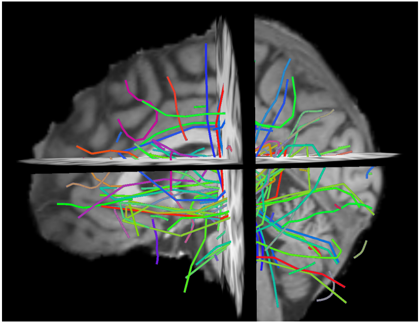
The Degree of Approximation: Pearson correlation

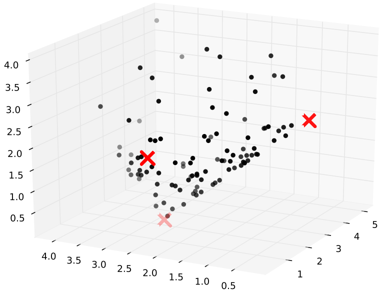
How to quantify the degree of approximation?

$$\mathbf{r}(d, \Delta_\Pi^d) = \frac{\sum_{X,X' \in S}(d(X,X') - \overline{d(X,X')})(\Delta_\Pi^d(X,X') - \overline{\Delta_\Pi^d(X,X')}))}{s_{d(X,X')} s_{\Delta_\Pi^d(X,X')}}$$

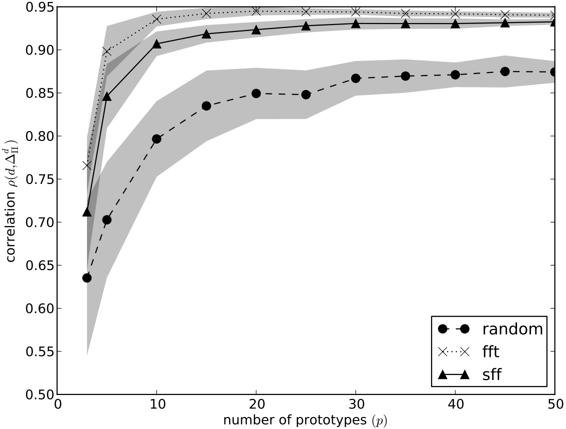
where $\Delta_\Pi^d(X,X') = ||\phi_\Pi^d(X) - \phi_\Pi^d(X')||_2$

Motivation: preserve relative distances (on average).





Experiment: Tractography data, 10^3 streamlines

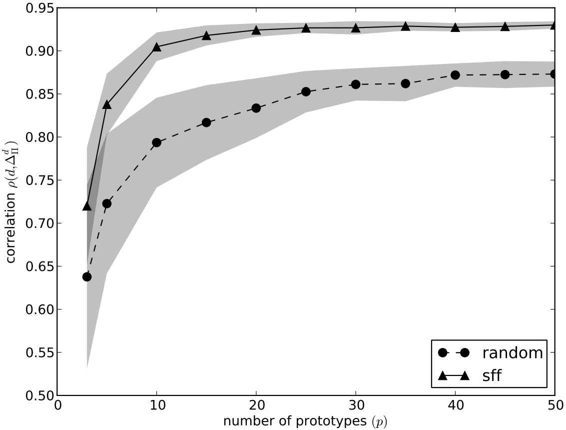


Timings ($p = 50$)

- FFT: < 1 secs. per iteration.
- SFF: 2 secs. per iteration.

50 iterations

Experiment: Tractography data, $\mathbf{3 \times 10^5}$ streamlines



Timings ($p = 50$)

- **FFT:** 15 **mins. per iteration. NOT COMPUTED**
- SFF: 2 secs. per iteration.

50 iterations

Conclusions & Future Work

Conclusions

- The *Dissimilarity Projection* with *SFF* is **accurate**: $r > 0.9$ with just $20 - 30$ prototypes.
- *Subset Farthest First* is **fast** on real tractographies.
- *Farthest First Traversal* is not advisable to embed tractographies.

Future Work

- Is correlation a good measure of approximation?
- Comparison against other Euclidean embeddings.

Thanks!

Garyfallidis, E. (2012).
Towards an accurate brain tractography.
PhD thesis, University of Cambridge.

Hochbaum, D. S. and Shmoys, D. B. (1985).
A Best Possible Heuristic for the k-Center Problem.
Mathematics of Operations Research, 10(2):180–184.

Olivetti, E. and Avesani, P. (2011).
Supervised segmentation of fiber tracts.
In *Proceedings of the First international conference on Similarity-based pattern recognition*, SIMBAD'11, pages 261–274, Berlin, Heidelberg. Springer-Verlag.

Pekalska, E., Duin, R., and Paclik, P. (2006).
Prototype selection for dissimilarity-based classifiers.
Pattern Recognition, 39(2):189–208.

Pekalska, E., Paclik, P., and Duin, R. P. W. (2002).

A generalized kernel approach to dissimilarity-based classification.
J. Mach. Learn. Res., 2:175–211.

Turnbull, D. and Elkan, C. (2005).
Fast recognition of musical genres using RBF networks.
Knowledge and Data Engineering, IEEE Transactions on, 17(4):580–584.

Zhang, S., Correia, S., and Laidlaw, D. H. (2008).
Identifying White-Matter Fiber Bundles in DTI Data Using an Automated Proximity-Based Fiber-Clustering Method.
Visualization and Computer Graphics, IEEE Transactions on, 14(5):1044–1053.