Evaluation of Euclidean Embeddings for Streamlines

Research Project Proposal

Kateryna Konotopska

March 16, 2018

The white matter of the brain can be represented as millions of streamlines by means of diffusion magnetic resonance imaging data and tractography algorithms. Streamlines are polylines with non-homogeneous number of points. Scalable algorithms to make inference from large datasets, e.g. clustering or spatial queries, usually require vectors. The Euclidean embedding of streamlines, i.e. how to accurately approximate streamline with vectors, is an open problem and current solutions are based on resampling or the dissimilarity representation. Exploring the distortion introduced by these and other Euclidean embeddings, e.g. the Lipschitz embeddings, is of primary interest in order to trade-off the accuracy of inferences and time of computation.

Given a finite metric space (S, d) with a distance metric d, where $S \subset U$ is a dataset of N objects drawn from a universe U, and $d : U \rightarrow \mathbb{R}^+$ is the original distance function on U, the mapping $F : S \rightarrow R^k$ with a distance metric d' is said to be an embedding. If d' is a Euclidean distance metric the embedding is called Euclidean embedding. A Euclidean embedding is said to have low distortion if $||F(a) - F(b)||_2 \approx d(a, b)$.

The project consists in developing and testing proposed embeddings over datasets of streamlines and evaluating their performance using different evaluation metrics. Possible metrics that can be used for evaluation are distortion, stress, time of execution and memory required.

As Euclidean embeddings, we consider the dissimilarity representation, the resampling of streamlines to a fixed of points and the Lipschitz embedding. The hosting laboratory can provide multiple datasets ready for these experiments. The code of the experiments will be developed on top of numerical libraries and neuroscientific libraries for the Python language.