Michela De Piccoli

A New Shape Similarity Framework for brain fibers classification

Ph.D. Thesis

July 9, 2018

Università degli Studi di Verona
Dipartimento di Informatica

Advisor:
Prof. Paolo Fiorini

Series N°: **TD-06-18**

Università di Verona
Dipartimento di Informatica
Strada le Grazie 15, 37134 Verona
Italy

"Correction does much, but encouragement does more."
-Johann Wolfgang von Goethe-

Summary

Diffusion Magnetic Resonance Imaging (dMRI) techniques provide a non-invasive way to explore organization and integrity of the white matter structures in human brain. dMRI quantifies in each voxel, the diffusion process of water molecules which are mechanically constrained in their motion by the axons of the neurons. This technique can be used in surgical planning and in the study of anatomical connectivity, brain changes and mental disorders. From dMRI data, white matter fiber tracts can be reconstructed using a class of technique called tractography. The dataset derived by tractography is composed by a large number of streamlines, which are sequences of points in 3D space. To simplify the visualization and analysis of white matter fiber tracts obtained from tracking algorithms, it is often necessary to group them into larger clusters or bundles. This step is called clustering. In order to perform clustering, a mathematical definition of fiber similarity (or more commonly a fiber distance) must be specified. On the basis of this metric, pairwise fiber distance can be computed and used as input for a clustering algorithm. The most common metrics used for distance measure are able to capture only the local relationship between streamlines but not the global structure of the fiber. The global structure refers to the variability of the shape. Together, local and global information, can define a better metric of similarity. We have extracted the global information using a mathematical representation based on the study of the tract with Frénet equations. In particular, we have defined some intrinsic parameters of the fibers that led to a classification of the tracts based on global geometrical characteristics. Using these parameters, a new distance metric for fiber similarity has been developed. For the evaluation of the goodness of the new metric, indices were used for a qualitative study of the results.

Acknowledgments

A sincere and esteemed thanks to Prof. Paolo Fiorini, not only for giving me the opportunity to undertake this path, but also for all that he has taught me on a human level. As he repeated to me several times: "The doctorate is a gym of life". And with so much training we managed to complete our work; one step at a time everything is done, but one must commit oneself and trust in what people say. I thank Prof. Gloria Menegaz and Prof. Carlo Combi, for their support and trust. Thanks to Prof. Arianna Menciassi and Dr. Elena De Momi for the revision of my thesis. Thanks to all the colleagues, even if the right term is Friends. You have been a great family in these years, you have supported me, consoled and given the splendid laughter. With our "hug time" you have taken away a thousand smiles even when things were not going well. In particular, I thank Marta for relaxing walks; Francesco for the return home full of complaints; David for all the advice; my "colleague" Fabrizio for having lightened the "sense of anxiety" of this last year. Thanks to Mario for bringing patience between nocturnal emails, deadlines and moments of despair ... it was a beautiful test of love. A special thanks goes to the people who have always supported my choices, my family. Without their support, they would have been much harsher years of security. Despite many times I did not have clear ideas, you never stopped believing in me. And then to you, that you have always been proud of any of my small steps ... your "my love you will make it" was my strong point ... I hope that from up there you are proud of me. I love you Papy.

Contents

List of Figures

1

Introduction

The human brain is the central organ of the human nervous system, and with the spinal cord makes up the central nervous system. It is the seat of recruitment, processing and transmission of information and the system of regulation of body functions. The brain directs the things we choose to do (like walking and talking) and the things our body does without thinking (like breathing). It is also in charge of sensing (sight, hearing, touch, taste, and smell), memory, emotions, and personality. Externally the brain is a soft, spongy mass of tissue. It is basically made up of two voluminous masses, the cerebral hemispheres, the surface full of pits and fissures that divide them into lobes. Internally a gray sub-stance and a white substance constitute it.

The gray matter is arranged peripherally and constitutes the cerebral cortex, while the white mat is in the center and consists of bundles of nerve fibers and of nuclei of gray matter, which are important nerve centers. The cerebral cortex is not functionally homogeneous, but divided into centers of localization, each one with different tasks. There is in fact a motor cortex, located in the frontal lobe, which processes and sends signals intended to produce the movements of the muscle groups; and a sensory cortex, which, placed in the parietal lobe, collects the stimuli from the periphery of our body.

With the use of brain imaging techniques, doctors and researchers have the opportunity to visualize behavior or clinical disorders within the human brain without recurring to invasive neurosurgery. There are a number of accepted, safe imaging techniques in use today in group facilities and hospitals throughout the world.

1.1 Neuroimaging

Neuroimaging or brain imaging is the use of various techniques to either directly or indirectly image the structure, function/pharmacology of the brain [24].

In 1918 the American neurosurgeon Walter Dandy introduced the technique of ventriculography [38]. X-ray [30] images of the ventricular system within the brain were obtained by injection of filtered air directly into one or both lateral ventricles of the brain. Dandy also observed that air introduced into the subarachnoid space via lumbar spinal puncture could enter the cerebral ventricles and also demonstrate the cerebrospinal fluid compartments around the base of the brain and over its surface. This technique was called pneumoencephalography. Moniz [7], a neurologist, accomplished the first cerebral arteriogram in 1927, whereby both normal and abnormal blood vessels in and around the brain could be visualized with great precision. In the early 1970s, Allan McLeod Cormack and Godfrey Newbold Hounsfield introduced computerized axial tomography (CAT or CT scanning) [29], and ever more detailed anatomic images of the brain became available for diagnostic and research purposes. Cormack and Hounsfield won the 1979 Nobel Prize for Physiology or Medicine for their work. Soon after the introduction of CAT in the early 1980s, the development of radioligands study of biomolecular behaviour, allowed single photon emission computed tomography (SPECT) [33] and positron emission tomography (PET) [9] of the brain. More or less concurrently, magnetic resonance imaging (MRI or MR scanning) [51] was developed by researchers including Peter Mansfield and Paul Lauterbur, who were awarded the Nobel Prize for Physiology or Medicine in 2003 [41]. In the early 1980s MRI was introduced clinically, and during the 1980s a veritable explosion of technical refinements and diagnostic MR applications took place. Scientists soon learned that the large blood flow changes measured by PET could also be imaged by the correct type of MRI. Functional magnetic resonance imaging (fMRI) [35] was born, and since the 1990s, fMRI has come to dominate the brain mapping field due to its low invasiveness, lack of radiation exposure, and relatively wide availability.

The ability to visualize anatomical connections between different parts of the brain, non-invasively and on an individual basis, has emerged as a major breakthrough for neuroscience's Human Connectome Project [1] . More recently, a new field has emerged, diffusion functional MRI (dfMRI) [42] as it was suggested that with DWI (Diffusion Weighted imaging) [42] one could also get images of neuronal activation in the brain. Finally, the method of diffusion MRI has also been shown to be sensitive to perfusion, as the movement of water in blood vessels mimics a random process, intra-voxel incoherent motion (IVIM) [43]. IVIM dMRI is rapidly becoming a major method to obtain images of perfusion in the body, especially for cancer detection and monitoring. In diffusion weighted imaging, the intensity of each image element (voxel) reflects the best estimate of the rate of water diffusion at that location. Because the mobility of water is driven by thermal agitation and highly dependent on its cellular environment, the hypothesis behind DWI is that findings may indicate (early) pathologic change. For instance, DWI is more sensitive to early changes after a stroke than more traditional MRI measurements such as T1 or T2 relaxation rates. A variant of diffusion weighted imaging, diffusion spectrum imaging (DSI) [72], was used in deriving the Connectome data sets; DSI is a variant of diffusion-weighted imaging that is sensitive to intra-voxel heterogeneities in diffusion directions caused by crossing fiber tracts and thus allows more accurate mapping of axonal trajectories than other diffusion imaging approaches. DWI is most applicable when the tissue of interest is dominated by isotropic water movement e.g. grey matter in the cerebral cortex and major brain nuclei, or in the body, where the diffusion rate appears to be the same when measured along any axis. However, DWI also remains sensitive to T1 and T2 relaxation. To entangle diffusion and relaxation effects on image contrast, one may obtain quantitative images of the diffusion coefficient, or more exactly the apparent diffusion coefficient (ADC) [44]. The ADC concept was introduced to take into account the fact that the diffusion process is complex in biological tissues and reflects several different mechanisms. Diffusion tensor imaging (DTI) [40] is important when a tissue such as the neural axons of white matter in the brain or muscle fibers in the heart has an internal fibrous structure analogous to the anisotropy of some crystals. Water will then diffuse more rapidly in the direction aligned with the internal structure, and more slowly as it moves perpendicular to the preferred direction. This also means that the measured rate of diffusion will differ depending on the direction from which an observer is looking. Traditionally, in diffusion-weighted imaging, three gradient-directions are applied, sufficient to estimate the different properties of diffusion tensor, for the measure of stroke. The principal direction of the diffusion tensor can be used to infer the white-matter connectivity of the brain (i.e. tractography; trying to see which part of the brain is connected to which other part). More extended DTI scans derive neural tract directional information from the data using 3D or multidimensional vector algorithms based on six or more gradient directions, sufficient to compute the diffusion tensor. The diffusion model is a rather simple model of the diffusion process, assuming homogeneity and linearity of the diffusion within each image voxel. Recently, more advanced models of the diffusion process have been proposed that aim to overcome the weaknesses of the diffusion tensor model. Amongst others, these include q-space imaging [39] and generalized diffusion tensor imaging.

1.1.1 Process of Diffusion

Diffusion is a passive form of material transport from areas of high concentration to areas of low concentration across a cell membrane. A distinguishing feature of diffusion is that it is dependent on particle random walk and results in mixing or mass transport, without requiring directed bulk motion.

The process of diffusion could be visualized by thinking of a drop of dark ink dropped into a glass of clear water. Initially, the ink appears to remain concentrated at the point of release. Gradually, some of the ink moves away from the region of high concentration, and instead of there being a dark region and a clear region, there is a graduation of color. After a long time, the ink is uniformly distributed in the water. The movement of the ink from the region of high concentration (ink drop) to the region of low concentration (the rest of the glass of water) is an illustration of the process of diffusion Figure(1.1).

Physis of Diffusion

There are two ways to describe the notion of diffusion: either a phenomenological approach starting with Fick's laws and their mathematical solutions, or a physical and atomistic one, by considering the random walk of the diffusing particles [58]. From the atomistic point of view, diffusion is considered as a result of the random walk of the diffusing particles. In molecular diffusion, the moving molecules are self-propelled by thermal energy. Random walk of small particles in suspension in a fluid was discovered in 1827 by Robert Brown. In 1826 a botanist Robert

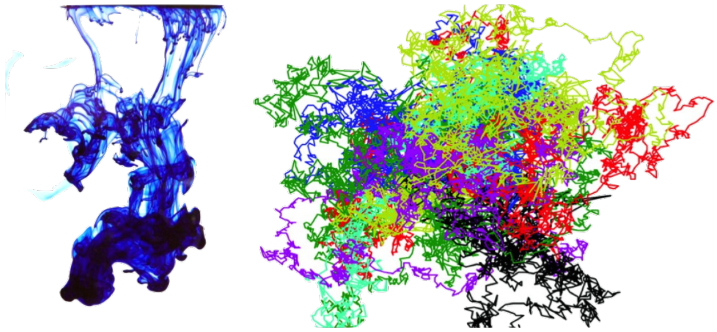


Fig. 1.1: Diffusion
Diffusion and Brownian Motion: on the left side of the figure, dark in into water can explain diffusion of ink in a fluid. On the right site of the picture trajectory of molecules show the random pathways of free diffusion.

Brown was studying the seemingly random pattern of motion that pollen grains exhibited when suspended in water through his microscope [37]. Initially puzzled, he attributed it to some biological phenomenon of the pollen, but when he later observed the same behavior with inanimate, inorganic substances, he rejected this hypothesis. It later became clear that the motion that he had observed was due to the buffeting of the pollen grains by water molecules surrounding them. This led to the revelation that liquids and gases were not static and lifeless as they might appear at first glance. The atoms and molecules from which they are constituent are in constant motion, undergoing persistent collisions and energy exchanges with other molecules and atoms. This phenomenon was called Brownian motion. The theory of the Brownian motion and the atomistic backgrounds of diffusion were developed by Albert Einstein [3].

The physical and mathematical theories of diffusion were developed and refined over the next two centuries by prominent scientists including Thomas Graham [61] and Adolf Fick [69], who developed a rigorous mathematical framework, which is still in use today Figure(1.2).

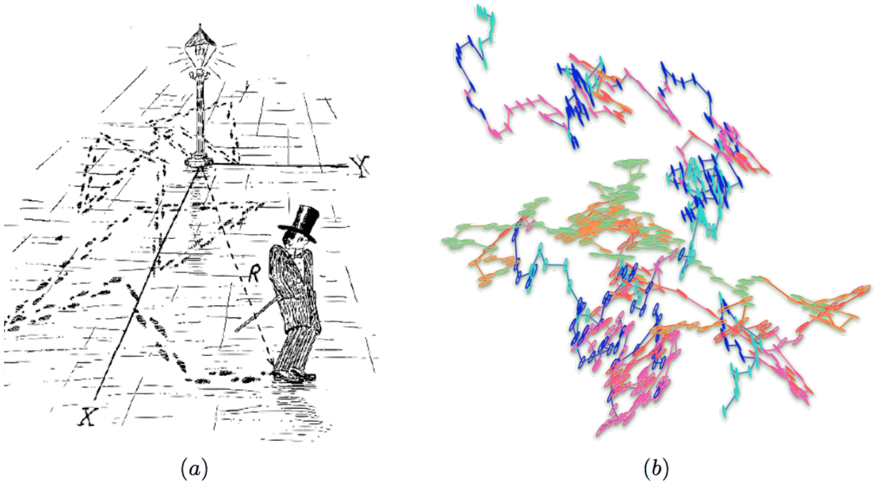
In 1855 Adolf Fick described and solved the diffusion coefficient D [16] by means of the first diffusion law:

$$J = -D\frac{d\varphi}{dx} \tag{1.1}$$

and the second diffusion law:

$$\frac{\partial\varphi}{\partial t} = D\frac{\partial^2\varphi}{\partial x^2} \tag{1.2}$$

where J is the flux, φ is the concentration, D is the proportional coefficient between flux and concentration, t was time-variable, x was the displacement variable, in the one dimensional case. In 1.1, Fick relates the diffusive flux to the concentration under the assumption of steady state. It postulates that the flux goes from regions of high concentration to regions of low concentration, with the concept that a solute



(a) (b)

Fig. 1.2: Random walk.
(a)Random walk according to Gamow; (b)Random walk trajectory due to Brownian motion in water.

will move from a region of high concentration to a region of low concentration across a concentration gradient. In 1.2, Fick predicts how diffusion causes the concentration to change with time.

In 1905 Albert Einstein, using Boltzmann's thermal energy predictions [50], derived a rule to estimate Avogadro's number by observing how far the polled grain moved over a given time. Einstein showed that the language of the Fick's Law still applied in the cases of self-diffusion where no macroscopic gradient existed. He provided a formulation of local probability to find the molecule and he realized a correlation between thermal fluctuations and diffusion coefficient. Einstein's explanation of Brownian motion was that the Brownian particles experience a net force resulting from the exterior collisions of surrounding water molecules. Einstein proved that the squared displacement of the particles from their starting point over a time, averaged over all of the sampled particles was directly proportional to the observation time; this formulation of diffusion model assumed that the medium was unrestricted and the particles therefore had equal mobility in every direction (free diffusion). The diffusion of particles inside fluids depends by the characteristics of the medium and obstacles characterize diffusion phenomena. Respect to the presence of preferable direction of molecules in their Brownian motion, we can define Figure(1.3):

- Anisotropic Diffusion, when a preferred direction is chosen by particles. This is typical in living tissues like fiber's muscles and cell's membranes. This means that obstacles characterize the medium as motion's constrictor.
- Isotropic Diffusion, when no direction is preferred by diffusion process. Ink in free water shows that the medium is isotropic. This means that no barriers obstacle the Brownian motion of particles.

- Restricted Diffusion occurs when particles are linked by membranes and barriers obstacle the diffusion process.
- Hindered Diffusion occurs when particles are linked by surrounding object but barriers don't obstacle diffusion propagation of molecules.

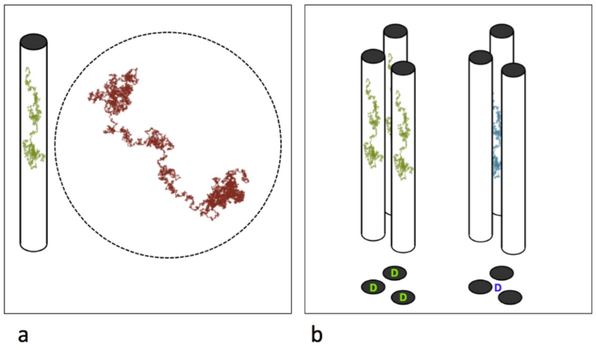


Fig. 1.3: Diffusion of particles
Picture a presents difference between anisotropic and isotropic diffusion (respectively left and right). Picture b show difference between restricted and hindered diffusion respectively left and right).

1.1.2 Diffusion MR Tractography

Diffusion MR imaging [56] of the brain was first adopted for use in clinical neuroradiology during the early 1990s and was found to have immediate utility for the evaluation of suspected acute ischemic stroke Figure(1.4).

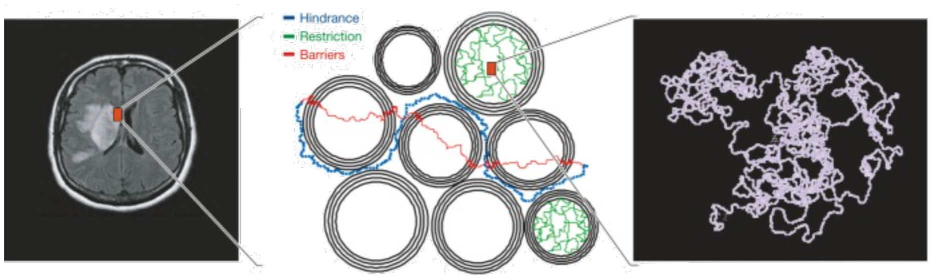


Fig. 1.4: Diffusion MR tractography
Diffusion-weighted image (left) reflect water diffusion behavior (random walk) (right). Diffusion behavior is modulated by tissue structure at the cellular level (middle).

Since that time, enormous strides forward in the technology of diffusion imaging have greatly improved image quality and enabled many new clinical applications.

These include the diagnosis of intracranial pyogenic infections, masses, trauma, and so on. In neuroscience, tractography is a 3D modeling technique used to visually represent neural tracts using data collected by difusion-weighted images. It uses special techniques of magnetic resonance imaging and computer-based image analysis. The results are presented in two-dimensional and three-dimensional images, Figure(1.5).

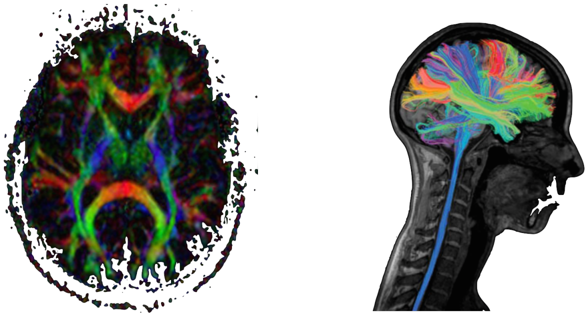


Fig. 1.5: MRI imaging
Results of MRI, in two-dimensional (left) and three-dimensional (right).

In addition to the long tracts that connect the brain to the rest of the body, there are complicated neural networks formed by short connections among different cortical and subcortical regions. The existence of these bundles has been revealed by histochemistry and biological techniques on post-mortem specimens. Brain tracts are not identifiable by direct exam, CT, or MRI scans. This difficulty explains the paucity of their description in neuroanatomy atlases and the poor understanding of their functions.

Tractography can be considered a large class of algorithms that generate a bundle interpretation of Diffusion Processes in the brain. This means that tractography is an interpretation of anatomy with a different grade of accuracy based on diffusion data elaborations. Fibers are obtained by following the paths of particles dropped in a vector field. The strategy used to approximate these paths constitutes the main difference between the methods analyzed.

White matter tractography algorithms can be classified into deterministic and probabilistic [8], [74]. The first only considers the main eigenvector direction in order to reconstruct the tract; the second class introduces the concept of perturbation in order to modify the vector direction at each location.

The deterministic approach produces a streamline that represents the main direction of diffusion for each voxel. This approach works well for many fiber bundles and can help to understand many configurations of lesion-pathways. The limit concerns the interpretation of voxels with low anisotropy. The Functional Anisotropy(FA) [10] interpretation is correct when voxels describe a region without any particular diffusion direction (ventricles). When a region is involved in fiber crossing, FA is around zero and the algorithm produces a wrong reconstruction (i.e. lateral portions of the corticospinal and corticobulbar tracts). The probabilistic approach aims to address this criticism by considering multiple pathways moving from the seed point and from each point along the reconstructed trajectories: the method accounts for the uncertainty in the estimation of fiber direction. Behind the probabilistic algorithm there are many pre-processing steps, such as co-registration, white matter extraction and statistical model estimation. This mean that it is very slow and non interactive.

Despite differences between strategies and algorithm implementation, the tractography procedure is based solving pathway equation [19]; common steps between different approaches are required to solve the pathway equation:

- Definition of seeds, the tracking process requires a starting point; this is typically known as seed region, a collection of voxel defined in derived tensor map. Drawing seed region requires anatomical knowledge of fiber bundles and their localization. Alternative approaches can used label definition of the brain from external sources: using digital label atlas and parcellation method, it is possible to automatically set seeds. This method can be imprecise and computationally expensive.
- Selection of an integration strategy: different integration methods can be applied in order to calculate pathways. The difference between them consists in the methodology of calculating the propagation direction and the step size. The most applied strategies are the Euler integration method, Runge-Kutta, Fiber Assignment by Continuous Tracking (FACT) and other alternative interpolation methods (Interpolated Streamline). As required for seed definition, stopping criteria determine when the tracking process has to finish. Thes methods can be defined as a combination of limits imposed by used definition that interrupts calculation of fiber tracts. Some example of it are FA threshold, Angle Threshold, maximal length of fiber and number of iterations.

Deterministic algorithms use a linear propagation approach. Fiber trajectories are generated in a stepwise fashion deriving the direction of each step from the local diffusion tensor. Approaches of deterministic category are the following:

- Streamline approaches [20] generate fibers following the direction of faster diffusion (often main eigenvector). The propagation direction is described by a linear propagation of diffusion measurements. The main drawback appeared in areas where diffusion propagation was not linear, such as planar regions, since the trace of the fiber can not be determined in order to partial volume effects, such as crossing, kissing, and branching. FACT (Fiber Assignment by Continuous Tracking) algorithm altered the propagation direction at the voxel boundary interfaces. FACT algorithm used variable step sizes,depending upon the length of the trajectory needed to pass through a voxel. The tensor deflection approach (TEND) was proposed in order to improve propagation in regions with low anisotropy, such as crossing fiber regions, where the direction of fastest diffusivity is not well defined. The idea is to use the entire D to deflect the incoming vector direction and to obtain a smoother tract reconstruction result.
- The tensor line propagation method incorporates information about the voxel orientation, as well as the anisotropic classification of the local tensor given anisotropic indices. Tensor line propagation direction is a combination of the

previous direction (main eigenvector) and the tensor deflection direction. Diffusion Toolkit and Trackvis are a famous couple of packages dedicated to diffusion tractography. The Diffusion Toolkit implements the most relevant deterministic algorithms such as the FACT, 2-order Runge Kutta, Interpolated Streamline and Tensorline. The Trackvis is the extraction software that calculates a sub-set of fiber bundle using ROI approaches.

Probabilistic fiber tractography algorithms aim to overcome bugs of deterministic approaches by adding some random choice. Probabilistic fiber tracking can be considered a simulation protocol that describes random walk of particles through a set of voxels guided by tensor rules. For each seed points, probability algorithms trace different paths and calculate the spatial probability distribution of connectivity. Final tracking is the spectrum of possible paths.

1.2 Fibers Tractography

In clinical application, it is common practice to focus on selected white matter fiber bundles; these bundles represent major pathways in the overall physical connectivity of the brain, i.e., groups of fibers belonging to the same anatomical regions Figure(1.6).

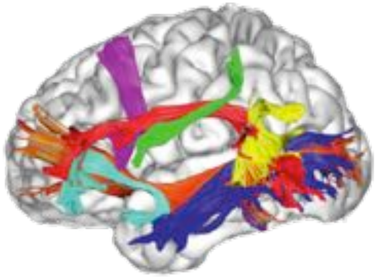


Fig. 1.6: Anatomical Bundles
This image represent some of anatomical fibers bundle in the brain.

Identify the structures that compose the different part of the brain is an important task in neuroimaging. One of the crucial points in neuroscience is image segmentation to partition the voxel into homogeneous subgroups that correspond to tissue types. Various techniques are proposed to segment fibers into meaningful anatomical structures for quantification and comparison between individuals.

Catani et al. [18] used a technique called virtual dissection to interactively select fibers passing through some manually defined regions of interests (ROIs). These approaches require a first manual intervention to select tracts of interest in a subset of subjects and then retrieve the same structure in other subjects, and making them unsuitable for a global WM segmentation. The manual identification of regions of interest is strongly affected by the prior knowledge used to identify the structures and very much prone to operator bias. Methods for the automatic decomposition of whole brain tractography into fiber bundles could greatly help reduce complexity and bias associated with manual segmentation. This approach, frequently referred to as tractography segmentation, aims at generating a simplified representation of the WM structure, enabling easier navigation and improved understanding of the structural organization of the brain and its overall connectivity. Though this technique is highly flexible, it is very time consuming due to a large amount of complex fiber structures. Moreover, the results may be biased by subjective opinions of experts.

Therefore, automatic fiber clustering algorithms, which require no user interaction and thus exclude undesirable bias have gained considerable attention. Clustering approaches represent a logical alternative to supervised methods as they permit to discover bundle structures without the need of prior anatomical knowledge. Using tractography technique, white matter fiber tracts are represented as streamlines, which are sequences of points in 3D space. These streamlines can be grouped into fiber bundles, i.e., groupings of streamlines with similar spatial and shape characteristics, based on anatomical knowledge using clustering methods.

In the tractography literature we can find approaches which use unsupervised and/or supervised learning algorithms to create bundles. In supervised learning the data sets are divided into training and a test set. For the training set, experts will have provided anatomical labels for a set of manually segmented streamline bundles. The task is then to identify similar structures amongst the unlabeled streamlines in the test set. In unsupervised learning the focus is on creating a partitioning of the streamlines without knowing any labels. Several fiber-clustering approaches have been described in the literature of which their main goal is to analyze a collection of tractography fiber paths in 3D and separate them into bundles, by taking advantage of their similarity. Most automatic techniques for segmenting fibers are based on geometric properties of fibers. Two fibers are usually grouped into a bundle if a small distance separates them, have comparable length and have similar shape. However, these criteria might be insufficient, since two fibers with different shapes can be grouped into a bundle if they start and end at the same region. So, fiber segmentation remains an area of active research and a fiber similarity measure is also needed to cluster fibers. A fiber similarity measure is a function that computes the (dis) similarity between pairs of fibers.

1.3 Aims of thesis

Recognize the anatomic bundles within the brain is a very important task in the scientific community. It is important both in the diagnostic phase, because a good display of the subdivision of the areas of the brain may help the surgeon in the general view of the problem and in the intraoperative phase the importance of identifying the fascicles relies on the repute for a good result in neurosurgical resection of the tumor, it is fundamental to preserve the areas of the main functional activities. Despite advances in the field of neuroimaging, this task is not easy to implement. There are many aspects to consider in order to get a good clustering algorithm that allows splitting the fibers of the brain in anatomical bundles.

Currently the surgeons, during surgery, make use of neuronavigation as a support tool. Neuronavigation represents the new frontier of minimally invasive surgery. A neuronavigation tool allows the surgeon to perform brain surgery with extreme accuracy and speed. Intraoperatively, the system shows on MRI or CT the tool position thus solving the first problem of the surgery, which is the spatial orientation of the anatomy. The possibility of preliminary planning the intervention and the help of a computer during the procedure create a new approach on how to consider the brain leading to innovative methods of surgical procedures. Through the use of neuron-navigation it is possible to identify the best path and the area where to attack the pathology considering its three-dimensional appearance and movements of adjacent structures.

As described above:

- From MR data, the white matter fiber tract can be reconstructed using a class of technique called tractography.
- The dataset derived by tractography is composed of a large number of streamlines, which are sequences of points in 3D space. To simplify the visualization and analysis of white matter fiber tracts obtained from MR data, it is often necessary to group them into larger clusters or bundles.
- In order to perform clustering, first a mathematical definition of fiber similarity (or more commonly a fiber distance) must be specified. Then, pairwise fiber distance may be calculated and used as input for a clustering algorithm.

The main problems that we are analyze:

- Define a relevant metric to assess the similarity of tracts.
- Establish the right number of clusters.
- Evaluate, once the clusters are obtained, qualitatively the results.

The final goal of this analysis is to obtain a new metrics for brain fiber classification. In literature various distance similarity measures are used; some methods calculates the similarity of the fiber point to point but these procedures are sensitive to noise, which affects the final distance similarity between pairs of fibers. The final distance alone between two fibers is not enough to tell whether they have a similar shape or they are separated by a small distance. Thus, this limits their ability to effectively group fibers into meaningful bundles. Another point is that the most common techniques have quadratic time complexity, is a problem for large fiber data sets. Several techniques are proposed to measure the shape similarity of two fibers; they also are efficient and robust to noise within fibers, but the complexity is still quadratic time.

Distance study during the last years captures the local relationship between pairwise fiber tracts but tend to lack the ability to captures the global structure of fiber tracts such as the shape variability and the neighborhood structures. In order to provide such information, the purpose of our proposed method, named "NewSimilarity" metric, is to developed a metric where both distance and shape information are considered during the clustering. The results obtained with our similarity metric confirm a good shape classification of the fibers of the brain.

In subsequent sections, we analyze the similarity metrics between fibers best known in the literature and we study the main clustering methods.

After this introduction, this thesis is organized as follows:

- *Chapter 2* : State Of the Art
  Overview of the distance metrics between the fibers and the clustering algorithms used in the literature.
  To simplify the visualization and analysis of white matter fiber tracts obtained from diffusion Magnetic Resonance Imaging data, it is often necessary to group fibers into larger clusters, or bundles. One approach is based on the knowledge of experts and is referred to as virtual dissection; it often requires several inclusion and exclusion regions of interest that make it a very hard process to reproduce across expert. Manual segmentation is very time consuming, is strongly user-dependent and adds biases to tract-based analyses.
  For this reason, it is preferable to use fiber clustering algorithms, which use distance metrics to classify the fibers in the different anatomical bundles.
- *Chapter 3* : Mathematical Background
  Overview of the mathematical foundations necessary in order to understand our method of studying the shape and classification of fibers.
  In particular we describe the properties to be verified by our metrics and how similarity metrics (or dissimilarity) are applied to clustering algorithms.
- *Chapter 4*: Method for our processing of create NewSimilarity metric
  New frameworks of clustering techniques and distance measures.
  The most common measures used for distance only capture the local relationship between streamlines but not the global structure of the fiber. Global structure, refer to the fiber variability shape. Together, local and global information, may define a good measure of similarity. In order to provide such information, we developed a novel framework where both distance and shape information are considered during the clustering.
- *Chapter 5* : Experimentals
  As a first step we consider datasets available online by ISMRM 2015 Tractography challenge-Data. Tractography challenge was based on an artificial phantom generated using the Fiberfox software, based on bundles segmented from a HCP subject.
- *Chapter 6* : Algorithm verification with clinical data
  In the following we show the results obtained by analyzing, as a first step, only the step related to the similarity metric. The clinical data used have been previously processed using the tractography algorithms and the bundles have already been segmented. The purpose is to understand which is the distance metrics that offers, compared to a ground truth, the best recognition of the different anatomical bundles. We also want to see if the metrics applied to online data are also applicable to clinical data. Afterwards we will summarize in a table the results obtained by applying the selected metrics to the clinical datasets and how clustering occurs in the different cases. We will present in detail the most significant cases. We use the library in Python, ?scikit-learn? for clustering classification.
- *Chapter 7* : Conclusions and future work

2

State of the Art

After pre-processing images and fit a diffusion model at every voxel, the fiber tracts can be virtually reconstructed or traced throughout the brain using computational methods called tractography. It is a method to reconstruct the pathways of major white matter fiber bundles, by fitting a curved path through the directional diffusion data at each voxel.

Deterministic tractography recovers fibers emanating from a seed voxel by following the principal direction of the diffusion tensor or the dominant direction of the diffusion orientation distribution function (ODF). However, this method has limitations: it depends on the choice of initial seed points and can be sensitive to the estimated principal directions. To overcome those drawbacks, probabilistic tractography methods have been proposed.

They can be computationally more intensive but can be more robust to partial volume averaging effects and uncertainties in the underlying fiber direction, which are inevitable due to imaging noise. Regardless of the chosen type, the analysis produces a large number of fibers that need to be grouped into anatomical bundles.

To simplify the visualization and analysis of white matter fiber tracts obtained from diffusion Magnetic Resonance Imaging data, it is often necessary to group them into larger clusters, or bundles. One way is based on the knowledge of experts and is referred to as virtual dissection; it often requires several inclusion and exclusion regions of interest that make it a process that is very hard to reproduce across expert. Manual segmentation is very time consuming, is strongly user-dependent and adds biases to tract-based analyses. For this reason, it is preferable to use fiber clustering algorithms, which use distance metrics to classify the fibers in the different anatomical bundles.

An overview of the distance metrics between the fibers and the clustering algorithms used in the literature is shown in Figure(2.1) where we have a general overview of the ingredients necessary for the study of a new similarity metric; in Figure(2.2) and Figure(2.3) we will see in detail, the similarity metrics and the clustering algorithms, used in the literature.

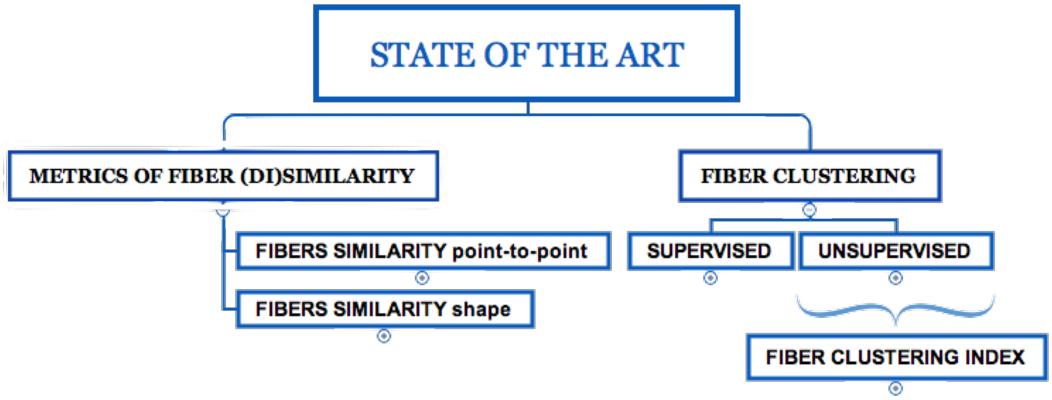


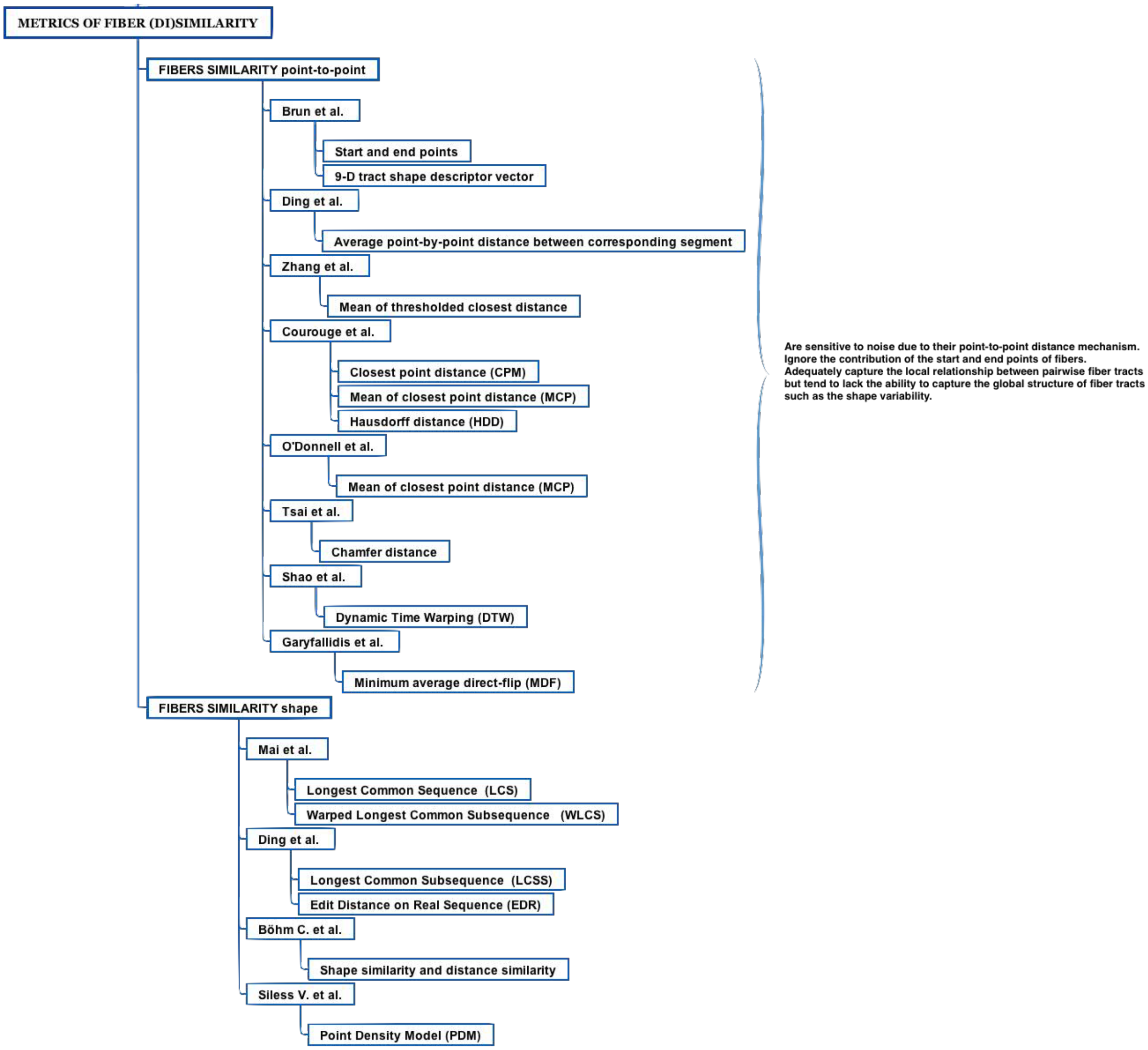
Fig. 2.1: State of the Art
Metrics of fiber (di)similarity and Fiber Clustering. Overview of the ingredients necessary for the study of a NewSimilarity metric.

2.1 Metrics of fiber (di)similarity: fibers similarity point-to-point and fibers similarity shape

A typical framework for fiber clustering defines a pairwise similarity/distance between each pair of fibers in a large set of candidate fibers, to group them into separate and distinct tracts. In order to perform clustering, first a fiber similarity measure must be specified, which is then used as input to a clustering algorithm. A fiber similarity measure is a function that computes the (dis)similarity between pairs of fibers. Two fibers are considered similar when they have comparable length, similar shape and are separated by a small distance.

Distances between pairwise fibers F_i and F_j are used for the similarity measure:

- Brun et al [5] considere the measure of similarity based on the idea that two fibers with similar end points should be considered similar. Euclidean distance between the end points of fibers is then used to calculate the fiber similarity. This similarity measure works fine in most cases where the fibers are not broken and really connect different parts of the brain in an anatomically correct way. This point is a great limitation because the white matter tracts contain many spurious and noisy fibers which do not allow an efficient calculus measure of similarity.
  The assumption done by Brun is not reasonable in many cases since not all fiber bundles start and end in the same regions.


METRICS OF FIBER (DI)SIMILARITY
FIBERS SIMILARITY point-to-point
Brun et al.
Start and end points
9-D tract shape descriptor vector
Ding et al.
Average point-by-point distance between corresponding segment
Zhang et al.
Mean of thresholded closest distance
Courouge et al.
Closest point distance (CPM)
Mean of closest point distance (MCP)
Hausdorff distance (HDD)
O'Donnell et al.
Mean of closest point distance (MCP)
Tsai et al.
Chamfer distance
Shao et al.
Dynamic Time Warping (DTW)
Garyfallidis et al.
Minimum average direct-flip (MDF)
FIBERS SIMILARITY shape
Mai et al.
Longest Common Sequence (LCS)
Warped Longest Common Subsequence (WLCS)
Ding et al.
Longest Common Subsequence (LCSS)
Edit Distance on Real Sequence (EDR)
Böhm C. et al.
Shape similarity and distance similarity
Siless V. et al.
Point Density Model (PDM)

Are sensitive to noise due to their point-to-point distance mechanism.
Ignore the contribution of the start and end points of fibers.
Adequately capture the local relationship between pairwise fiber tracts but tend to lack the ability to capture the global structure of fiber tracts such as the shape variability.

Fig. 2.2: State of the Art
Metrics of fiber (di)similarity. Review of principal distance used in literature.

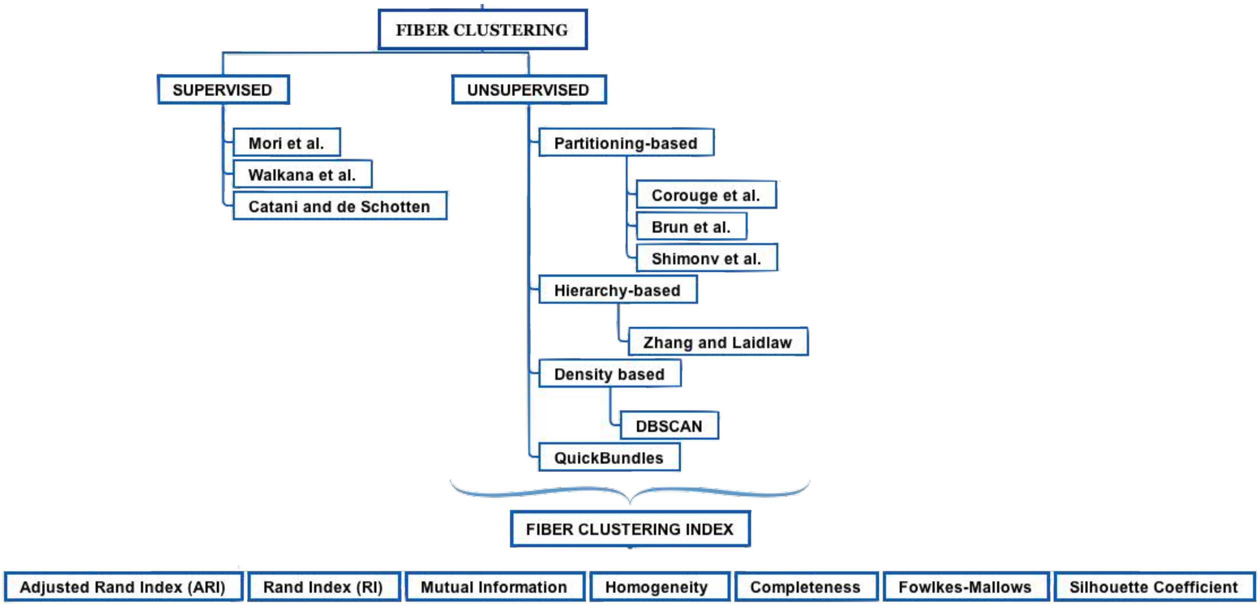


Fig. 2.3: State of the Art
Fiber Clustering and Index. Review of principal clustering algorithms used in literature.

- Later Brun et al. introduced a 9-D tract shape descriptor vector, defined as the mean and lower triangular part of the covariance matrix of the points on a fiber, and computed the Euclidean distance between shape descriptors [4]. It ignores the information of most other points and the fiber pairwise shape similarity.
- Corouge et al [32] form point pairs by mapping each point of one fiber to the closest point on the other fiber; the distance between the pairs of fibers is calculated using these points. Three distances are defined; the first one is the *closest point distance*:

$$\begin{gathered} d_c(F_i, Fj) = min_{pk\in F_i, p_l\in F_j} ||p_k - p_l|| \\ \text{with} \quad ||\cdot|| \quad \text{being the Euclidean norm} \end{gathered} \tag{2.1}$$

It is the minimum distance between a pair of points. These distance cannot be expected to have a good discrimination power between fibers since it encodes only very raw information about fiber similarity and closeness.
The second distance is the *mean of closest point distances (MCP)*:

$$\begin{gathered} d_M(F_i, Fj) = mean(d_m(F_i, F_j), d_m(F_i, F_j)) \\ \text{with} \quad d_m(F_i, F_j) = mean_{p_l\in F_i} min_{p_k\in F_j} ||p_k - p_l|| \end{gathered} \tag{2.2}$$

which calculates the average of the points pair distances. The minimum distance provides a global similarity measure integrated along the whole curve.

The last measure is the *Hausdorff distance (HDD)*:

$$\begin{aligned} d_H(F_i, F_j) &= max(d_h(F_i, F_j), d_h(F_i, F_j))) \\ \text{with} \quad d_h(F_i, F_j) &= max_{p_k \in F_i} min_{p_l \in F_j} ||p_k - p_l|| \end{aligned} \tag{2.3}$$

that is the maximum distance between a pair of points.
Being a worst-case distance, the Hausdorff distance is a useful metric to reject outliers and prevents the algorithm from clustering curves with high dissimilarity.

- Zhang and Laidlaw [76] define the mean of thresholded closest distances; the distance between two fibers as the average distance from any point on the shorter fiber to the closest point on the longer fiber, and only distances above a certain threshold contribute to this average.
- The average point-by-point distance between corresponding segment is defined by Ding et al. [22]. Their fiber similarity measure is then defined as the mean distance between the corresponding segments.
  When two fibers are similar in shape and close in location these distance is small; when the distance between a pair of fibers is large, or their shapes are different, the distance is large. To define piece-wise similarity, the authors use the mean Euclidean distance between the segments. This similarity method is efficient but not effective since this measure also loses the point-by-point information.
- The distance measure such as closest point distance, Chamfer distance [26]:

$$\begin{aligned} D(F_i, F_j) = \frac{1}{|F_i|} \sum_{F_{i_p} \in F_i} & min_{F_{j_q} \in F_j} ||F_{i_p} - F_{j_q}|| \\ & \text{with} \quad F_i = F_{i_p} \quad p = 1, \quad F_j = F_{j_q} \quad q = 1 \\ & \text{where p,q are the set of points on a fiber } F_i, F_j \\ & || \cdot || \quad \text{Euclidean norm} \end{aligned} \tag{2.4}$$

  and Hausdorff distance [31], have been the most used that adequately capture the local relationship between pairwise fiber tracts.
  In [46] the authors propose a novel approach for measuring the similarity of 3D curves in a large dataset that includes the whole information of the curve for more accurate clustering and further quantitative analysis.
  In [6] a manifold learning approach to fiber tract clustering is using. To generate the similarity measure, the Chamfer and Hausdorff distance are initially employed as a local distance metric to construct minimum spanning trees between pairwise fiber tracts.
  All these distances lack the ability to capture the global structure of fiber tracts such as the shape variability and the neighborhood structures. In [53] several distance measures are implemented for fiber evaluation and the authors concluded that the mean of closest distances performs better than closest point distance, Hausdorff distance and end point distance.
- Shao et al. [34] extended the idea of the functions proposed by Courouge, using Dynamic Time Warping (DTW):

$$d_{dtw}(F_i, F_j) = ||p_n - q_m|| + min(d_{dtw}(f_{i_{n-1}}, f_{j_m}), d_{dtw}(f_{i_n}, f_{j_{m-1}}), d_{dtw}(f_{i_{n-1}}, f_{j_{m-1}}))$$
$$\text{with} \quad F_i = (p_1, ..., p_n), F_j = (q_1, ..., q_m) \tag{2.5}$$

due to its flexibility with varying length fibers. It is a technique that looks for the optimal alignment of two time series.

- Garyfallidis et al. [23] used Minimum average Direct-Flip (MDF) distance:

$$\begin{aligned} MDF(F,t) &= min(d_{direct}(F,t), d_{flipped}(F,t)) \\ d_{direct}(F,t) &= d(F,t) = \frac{1}{K}\sum_{K} i = 1|F_i - t_i| \\ d_{flipped}(F,t) &= d(F, t^{flip}) = d(F^{flip}, t) \end{aligned} \tag{2.6}$$

  which is a symmetric distance function that can address with the streamline bi-directionality problem. The direct distance $d_{direct}(F,t)$ between two streamlines F,t is the mean of the Euclidean distances between corresponding points. MDF can be applied only when both streamlines have the same number of points.

  The main advantages of the MDF distance are that it is fast to compute, it takes account of streamline direction issues through consideration of both direct and flipped streamlines.

The distance measures similar to those mentioned above:

- Are sensitive to noise due to their point-to-point distance mechanism which affects the final distance similarity between pairs of fibers.
- Ignore the contribution of the start and end points of fibers which play an important role in the segmentation.
- Adequately capture the local relationship between pairwise fiber tracts but tend to lack the ability to capture the global structure of fiber tracts such as the shape variability and the neighborhood structures, due to the complexity of fiber structure.
- Their quadratic time complexity makes them hard to deal with large fiber datasets.

For these reason other approaches are being studied. These metrics, compared to previous metrics, are better to an analysis of fibers regarding the shape.

Two key aspects for achieving effectiveness and efficiency when managing time series data are representation methods and similarity measures. Time series are essentially high dimensional data and directly dealing with such data in its raw format is very expensive in terms of processing and memory cost. The best known of such distance is the Longest Common Subsequence (LCSS) distance [47]:

$$LCSS(MBE_{\delta,\varepsilon}(A), B) = \sum_{i=1}^{n} \tag{2.7}$$

that utilizes the longest common subsequence model. Given a time constrain δ and a similarity threshold ε, the lower bounding distance of $LCSS_{\delta,\varepsilon}(A,B)$ of two equal length fibers can be calculated by using the Minimum Bounding Envelope of

$A(MBE_{\delta,\varepsilon}(A))$ with respect to the time constraint δ and the similarity threshold ε.

To adapt the concept of matching characters in the settings of time series, a threshold parameter was introduced, starting that two different points from two time series are considered to match if their distance is less than the parameter.

Edit Distance on Real sequence (EDR) [27], given two trajectories R and S of lengths n and m, respectively the distance between R and S is the number of insert, delete, or replace operations that are needed to change R into S,

$$EDR_{\varepsilon}(R,S) = \frac{edr_{\varepsilon}(R,S)}{max(n,m)} \tag{2.8}$$

Similar to LCSS, EDR also uses a threshold parameter, except its role is to quantify the distance between a pair of points to 0 or 1.

- Mai et al. [67] introduce a novel similarity model called SIM for fiber segmentation, using a so-called fiber envelope. Based on this scheme, some new shape similarity techniques are proposed by adapting existing similarity techniques for trajectory data such as those mentioned before, Warped Longest Common Subsequence (WLCS) is one of these.
  Given a time constrain δ and a similarity threshold ε, the similarity between two sequences A, B, with time points n, m is:

  $$WLC_{\delta,\varepsilon}(A,B) = 1 - \frac{wlcs_{\delta,\varepsilon(A,B)}}{n,m} \tag{2.9}$$

  is more accurate and more robust to noise and local time shifting within fibers than other similarity measures.
- Böhm C. et al [15], propose a novel similarity measure for fiber clustering by combining shape similarity and distance similarity into a unified and flexible method.
  LCSS is specially adapted to points in three-dimensional space to measure shape similarity, and is less sensitive to noise than other distance based methods. In addition, the distance between start and end points of a pair of fibers, which is referred to as distance similarity, is also incorporated to effectively capture the complex notion of fiber similarity.
  As a result, our approach provides an effective and flexible way to the similarity between fibers.
- Siless V. et al [64], propose tu use the Point Density Model (PDM) metric. Given a fiber X, it is represented as the sum of Dirac concentrated at each fiber point, respect to fiber Y:

  $$PDM^2(X,Y) = ||X||^2 + ||Y||^2 - 2\langle X,Y\rangle \tag{2.10}$$

Whereas PDM time complexity is quadratic in the number of points per fiber, using it for computing a full distance matrix is too expensive time-wise. By using multidimensional scaling the authors only compute a partial distance matrix and embed this information in a new set of fiber-like points. PDM is sensitive to the fibers form and position and is quite robust to missing fiber segments.

This last property is much desired as fibers are often mis-segmented due to noise and crossing fibers issues. MDF distance and PDM distance require us to resample tracts so that they have the same amount of points each. Hausdorff and Mean Closest Point work for tracts with different amounts of points.

2.2 Fiber Classification

Diffusion-Weighted Magnetic Resonance Imaging is one of the most used techniques for the analysis of the human brain white matter. Tractography datasets, composed of a big set of 3D streamlines, can be reconstructed from dMRI and represent the main anatomical connections in the brain.

There are two ways for analysis of dMRI tractography data generatting a quantitative description of the white matter connections. The first method, fiber clustering, describes the connections of the white matter as clusters of fiber trajectories. The clusters give anatomical regions in which properties of the white matter structure may be measured.

The second method is parcellation-based and uses tractography to estimate the "structural connectivity" between pairs of parcellated cortical regions. The pairwise connectivities are encoded in a matrix that models networks in the brain.

These two types of analysis of dMRI tractography data both perform a segmentation of the white matter, but with different goals.

Parcellation-based approaches for white matter segmentation address the question of what regions a fiber trajectory may connect. These approaches take advantage of additional information in the form of a cortical parcellation into regions of interest (ROIs) that define network nodes, enabling analysis of the brain as a network. Once the nodes are defined, segmentation of tractography is straightforward and is based simply on connections between ROIs.

Clustering approaches generally address the goal of detecting the central, anatomically named portions of each fiber tract, without reference to cortical regions. These approaches do not enable graphical analysis of the brain networks, but rather focus on measuring properties of the anatomy of the fiber tracts. Early work in tractography clustering had the goal of organizing the fibers within a single subject into fiber tracts or bundles.

In the literature this problem is divided into two parts:

- choice of similarity or distance metric for comparing fibers (methods explained in the previous section)
- choice of clustering method

In the following section, we especially focus on reviewing the fiber clustering methods, which is the scope of our work.

2.2.1 Fiber Clustering: Supervised and Unsupervised

In a supervised setting, the classes are a predetermined finite set. A learning data set is labeled with the classifications. The task of the algorithm is to find predictive patterns and build mathematical models to relate those to the known classification.

Unsupervised algorithms do not start with a classification; they search for similarities in the data to determine if they can be characterized as belonging to the same group. In an unsupervised setting, or "cluster analysis", the algorithm does not rely on any information on how elements are grouped and its task is to group them. Several methods have been proposed to obtain the clear fiber bundles representation.

We analyze as first step the Supervised methods: the identification of fiber bundles is carried out via manual identification of regions of interest corresponding to the main known pathways. The three mains algorithms of this type are:

- Mori et al. [54]
- Wakana et al. [71]
- Catani et al. [17]

However, these technique requires a priori knowledge about the trajectory and can be used only for well-characterized white matter tracts.

To reduce the complexity of manual segmentation and reduce the errors introduced by it, methods are introduced for the automatic decomposition of whole brain tractography into fiber bundles. For this reason over the years, many semi-automatic methods have been studied in the scientific community for determining the bundles within and across subjects with little or no human intervention.

This approach, frequently referred to as tractography segmentation, aims at generating a simplified representation of the white matter structure, enabling easier navigation and improved understanding of the structural organization of the brain and its overall connectivity.

To automate bundles retrieval, various methods, based on different strategy were proposed over the last few years. The solution proposed in Li et al. [45] is an evolution of the ROI-based technique that works directly on fiber and applies prior knowledge to perform preliminary parcellation of the brain. A nonlinear method of kernel-principal component analysis (PCA) is used to project the fiber curves onto the principal component space of the kernel vectors.

Then, a fuzzy c-mean algorithm is applied to automatically group the fibers in the feature space. However, this approach is limited by the level of detail of the brain atlases, which can prevent the retrieval of small structures or suffer from cross-subject misalignments.

Other methods were also proposed to recover local white matter bundles using prior knowledge. Mayer et al. [49] presented a supervised framework for the automatic registration and segmentation of white matter tractography extracted from brain diffusion magnetic resonance imaging.

The framework relies on the direct registration between the fibers, without requiring any intensity-based registration as preprocessing. An affine transform is recovered together with a set of segmented fibers. These approaches require a first manual intervention to select tracts of interest in a subset of subjects and then retrieve the same structure in other subjects, and making them unsuitable for a global white matter segmentation.

The second method for fiber classification is "Unsupervised". Indeed clustering approaches represent a logical alternative to supervised methods as they permit to discover bundle structures without the need of prior anatomical knowledge.

In unsupervised method, we can classify the clustering techniques on the basis of the type of algorithm used to divide space:

- 1 Partitioning methods.
  A partitioning method first creates an initial set of k partitions, where, parameter k is the number of partitions to construct. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. These clustering techniques create a one-level partitioning of the data points.
  The best known algorithms in this area are K-means, K-medoids and fuzzy C-means. The clusters are formed according to the distance between data points and cluster centers are formed for each cluster. The number of clusters (k-value) is specified by the user. The data points in each cluster are displayed by different colors, one color for one cluster.
  K-Means is one of the simplest unsupervised learning algorithms that solves the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of cluster fixed a priori.
  The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.
  The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When, no point is pending, the first step is completed and an early group age is done. At this point it is necessary to re-calculate k new centroids as bar centers of the clusters resulting from the previous step.
  After obtaining these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid.
  A loop has been generated. As a result of this loop, one may notice that the k centroids change their location step by step until no more changes are done. K-means is a simple algorithm that has been adapted to many problem domains and it is a good candidate to work for a randomly generated data points.
  One of the most popular heuristics for solving the K-means problem is based on a simple iterative scheme for finding a locally minimal solution.
  The basic strategy of K-medoids clustering algorithms is to find k clusters in n objects by first arbitrarily finding a representative object (the medoids) for each cluster.
  Each remaining object is clustered with the medoid to which it is the most similar. K-medoids method uses representative objects as reference points instead of taking the mean value of the objects in each cluster.
  The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects.
  Traditional clustering approaches generate partitions; in a partition, each pattern belongs to one and only one cluster.
  Hence, the clusters in a hard clustering are disjoint. Fuzzy clustering extends this notion to associate each pattern with every cluster using a membership function.

The output of such algorithms is a clustering, but not a partition. Fuzzy clustering is a widely applied method for obtaining fuzzy models from data. It has been applied successfully in various fields including geographical surveying, finance or marketing.

K-means algorithm is efficient for smaller data sets and K-medoids algorithm seems to perform better for large data sets.

The performance of Fuzzy clustering is intermediary between them. It produces close results to K-means clustering, yet it requires more computation time than K-means because of the fuzzy measures calculations involved in the algorithm.

- 2 Hierarchical methods.

  A partitional clustering is a simply a division of the set of data objects into non-overlapping subsets such that each data object is in exactly one subset.

  A hierarchical clustering is a set of nested clusters that are organized as a tree. Hierarchical clustering algorithm is of two types, Agglomerative Hierarchical clustering algorithm and Divisive Hierarchical clustering algorithm; both this algorithm are exactly reverse of each other.

  Agglomerative Hierarchical clustering works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. Again distance between the data point is recalculated but which distance to consider when the groups has been formed. There are many available methods, some of them are single linkage, complete linkage, average linkage and centroid distance.

  The advantages of this method is no a-priori information about the number of clusters required and are easy to implement and gives best result in some cases.

  As disadvantages, the algorithm can never undo what was done previously, time complexity, sensitivity to noise and outliers and breaking large clusters.

- 3 Density-Based.

  In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas are usually considered to be noise and border points.

  The most popular density based clustering method is DBSCAN. In contrast to many newer methods, it features a well-defined cluster model called "density-reachability". Similar to linkage based clustering, it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius.

  A cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects' range.

  Another interesting property of DBSCAN is that its complexity is fairly low - it requires a linear number of range queries on the database - and that it will discover essentially the same results (it is deterministic for core and noise points, but not for border points) in each run, therefore there is no need to run it multiple times.

  Mean-shift is a clustering approach where each object is moved to the densest area in its vicinity, based on kernel density estimation. Eventually, objects con-

verge to local maxima of density. Similar to k-means clustering, these "density attractors" can serve as representatives for the data set, but mean-shift can detect arbitrary-shaped clusters similar to DBSCAN.
Due to the expensive iterative procedure and density estimation, mean-shift is usually slower than DBSCAN or k-Means. Besides that, the applicability of the mean-shift algorithm to multidimensional data is hindered by the unsmooth behaviour of the kernel density estimate, which results in over-fragmentation of cluster tails.

A common clustering framework is based on the exploitation of the affinity matrix of a single subject that indicates the similarity between each pair of fibers:

- Brun et al. [4] present a framework for unsupervised segmentation of white matter fiber traces obtained from diffusion weighted MRI data.
  Fibers are compared pairwise to create a weighted undirected graph which is partitioned into coherent sets using the normalized cut criterion. The points to be clustered are represented by a undirected graph, where the nodes correspond to the points to be clustered and each edge represent the similarity between points. The cut is a graph theoretical concept which for a partition of the nodes into two disjunct sets.
- O'Donnell et al. [55] use the eigenvectors of the similarity matrix. An eigenvector is a vector that when multiplied by the matrix, still points in the same direction.
  The authors used the top eigenvectors of the fiber similarity matrix to calculate the most important shape similarity information for each fiber path while removing noise. For each fiber path, this information could be visualized as a point, to show the separation of clusters according to similarity.
- Zhang et al. [75] used the agglomerative hierarchical clustering method and defined the distance between two clusters as the minimum proximity value between any two curves from two clusters.

A limitation common to all algorithms based on affinity matrix is their propensity to suffer from computational load owing to the calculation of pairwise distances between streamlines.

Approaches to reduce computational complexity have been proposed like:

- Quick Bundles (QB) [23]. This method overcomes the complexity of these large data sets and provides informative clusters in seconds.
  Each cluster can be represented by a single centroid streamline, collectively these centroid streamlines can be taken as an effective representation of the tractography. QB is a surprisingly simple and fast algorithm which can reduce tractography representation to an accessible structure in a time that is linear in the number of streamlines.
  The streamlines, are a fixed-length ordered sequence of points.
- Demir et al. [21] adopt the scenario of clustering data streams into the fiber clustering framework.
  Existing clustering methods often suffer from the burden of computing pairwise fiber (dis)similarities, which escalates quadratically as the number of fiber pathways increases. To address this challenge, the authors propose to use an

online hierarchical clustering method, which yields a framework similar to doing clustering while simultaneously performing tractography.

- Ros et al. [60] uses a hierarchical cluster analysis approach that exploits the inherent redundancy in large datasets to time-efficiently group fiber tracts. Structural information of a white matter atlas can be incorporated into the clustering to achieve an anatomically correct and reproducible grouping of fiber tracts.

  This approach facilitates not only the identification of the bundles corresponding to the classes of the atlas; it also enables the extraction of bundles that are not present in the atlas.
- In Guevara et al. [25], the authors describe a sequence of algorithms performing a robust hierarchical clustering that can deal with millions of diffusion-based tracts. The end result is a set of a few thousand homogeneous bundles. This simplified representation of white matter that can be used further for group analysis.

  The bundles can also be labelled using ROI-based strategies in order to perform bundle oriented morphometry.

The methods described above, have been proposals to automatically estimate the number of cluster from the dataset. However, the results of this approach are strongly conditioned by the number of hierarchical steps and several input parameters are required to carry out a comprehensive map of white matter bundles.

2.3 Validation

Clinical applications of diffusion MR fibre tractography tend to be more concerned with the anatomical trajectory of the fibres rather than structural or functional connectivity, i.e. to determine the exact spatial location of white matter pathways in the human brain.

For clinical application it is therefore not only important to validate what white matter pathways connect to what brain region, but also validate their exact course through the brain. In following, we will discuss of methods available for validation of the precise anatomical trajectory of diffusion MR fibre tractography and we address their benefits and drawbacks.

One more challenges still remaining in validating tractography is defining a proper evaluation metric for comparing the tractography results with the established ground truth. Fibre tractography in its current state is inherently limited by the spatial resolution of MRI and noise.

Tractography results are therefore unlikely to exactly coincide with the true anatomy. If we want to compare the performance of fibre tractography algorithms, we not only need to establish a ground truth, but also require a metric to compare the performance. A common metric in image analysis with respect to segmentation is the use of volume overlap to determine how well two segmentations coincide.

One common approach to validating the accuracy and performance of clinical scanners of any kind is the use of physical phantoms.

Two distinct classes of physical phantoms are available.

The first class of phantoms are the phantoms constructed from artificial fibrous materials such a polymer or yarn.

The second class of phantoms are the biological tissue samples. Phantoms are constructed from materials with known properties and modelled in such a way it represents the part of the body of interest. Finding the right materials and constructing a descent model are the major challenges for physical phantoms. The phantoms tend to be overly simplistic compared to the complex structure of the brain, although an attempt has been made to create a semi-realistic phantom for use in the FiberCup competition.

This semi-realistic phantom contains all forms of white matter tracts that are common to the human brain. Another limitation to physical phantoms is that it can be difficult to select the right material to match properties of white matter. Therefore any polymer material would face the same limitations; this limitation has been overcome by using biological tissue. Another validation technique that belongs to the phantom class is software models. Based on the models on white matter structures from neuroanatomy that are available in large-scale brain maps and atlases, artificial diffusion-weighted MR images can be synthesised. The main advantage of software simulations is that any shape imaginable can be modelled using synthetic tensor fields.

Moreover, the imaging parameters can be set to any preferred value and the exact position of the tracts in known with high-precision. No other class of phantoms or validation technique allows this much control over the parameters. Additional benefits are achieved when comparing the derived results with the ground truth as no registration of the two images is requires since they operate within the same coordinate space.

A disadvantage of this technique is that generation of a realistic model is computationally intensive.

2.3.1 Clastering Results Validation

The problem of evaluating models in unsupervised settings is notoriously difficult; evaluating the performance of a clustering algorithm is not as trivial as counting the number of errors or the precision and recall of a supervised classification algorithm.

In particular any evaluation metric should not take the absolute values of the cluster labels into account but rather if this clustering defines separations of the data similar to some ground truth set of classes or satisfying some assumption such that members belong to the same class are more similar that members of different classes according to some similarity metric.

The process of clustering validation that we utilizing for our datasets is summarize in Figure(2.4). We considering the final result of tractography algorithm, then the bundles fibers; these fibers are extracted by expert clinician and utilized it as ground truth. Cluster of these fibers is formed by with two ingredients, the first is the measure of similarity between them, the second the clustering algorithm. This step produce a classification of all fibers where their evaluation is generating by studying of index of measures of cluster.

For clustering metric validation, there is one set of standard criteria, given the knowledge of the ground truth:

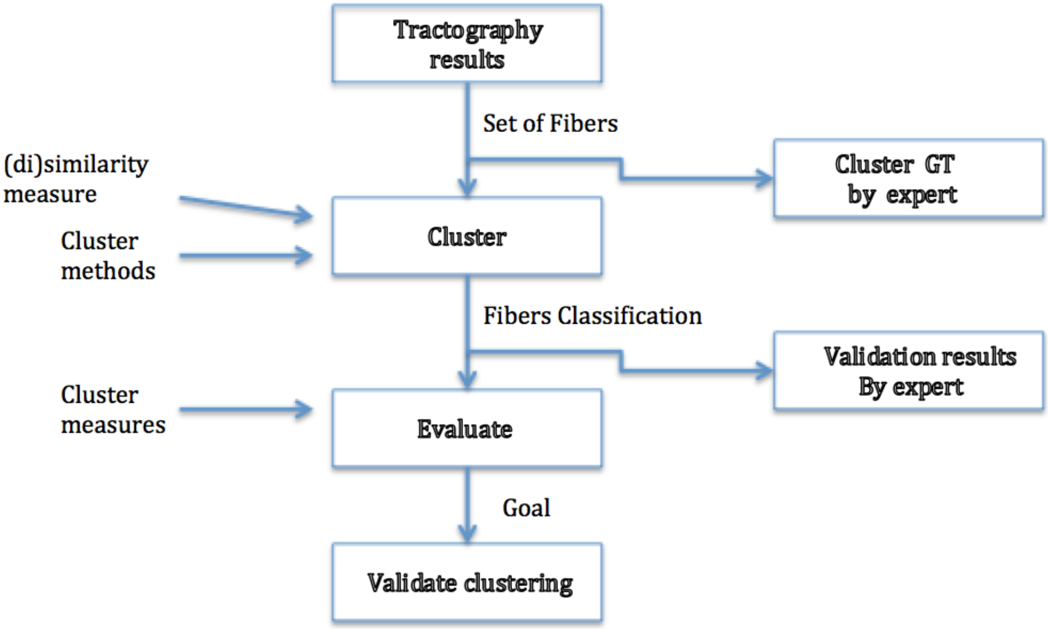


Fig. 2.4: Tractography results
Scheme of clustering validation.

- The Adjusted Rand Index (ARI) computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. If C is a ground truth class assignment and K the clustering, let us define a and b as:
  - a, the number of pairs of elements that are in the same set in C and in the same set in K
  - b, the number of pairs of elements that are in different sets in C and in different sets in K

  The raw (unadjusted) Rand index is then given by:

  $$RI = \frac{a+b}{C_2^{n_{samples}}} \tag{2.11}$$

  Where $C_2^{n_{samples}}$ is the total number of possible pairs in the dataset (without ordering).
  Advantages:
  - random (uniform) label assignments have a ARI score close to 0.0 for any value of $n_{clusters}$ and $n_{samples}$
  - bounded range $[-1, 1]$: negative values are bad (independent labelings), similar clusterings have a positive ARI, 1.0 is the perfect match score
  - no assumption is made on the cluster structure: can be used to compare clustering algorithms such as k-means

- Mutual Information is a function that measures the agreement of the two assignments, ignoring permutations.
  Two different normalized versions of this measure are available, Normalized Mutual Information(NMI) and Adjusted Mutual Information(AMI).
  Assume two label assignments (of the same N objects), U and V. Their entropy is the amount of uncertainty for a partition set, defined by:

  $$H(U) = -\sum i = 1|U|P(i)log(P(i)) \tag{2.12}$$

  where $P(i) = |U_i|/N$ is the probability that an object picked at random from U falls into class U_i. Likewise for V:

  $$H(V) = -\sum j = 1|V|P^{'}(j)log(P^{'}(j)) \tag{2.13}$$

  with $P^{'}(j) = |V_j|/N$. The mutual information (MI) between U and V is calculated by:

  $$MI(U,V) = \sum i = 1|U|\sum j = 1|V|P(i,j)log(\frac{P(i,j)}{P(i)P^{'}(j)}) \tag{2.14}$$

  where $P(i,j) = |U_i \cup V_j|/N$ is the probability that an object picked at random falls into both classes U_i and V_j. It also can be expressed in set cardinality formulation:

  $$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|}\frac{|U_i \cap V_j|}{N}\log\left(\frac{N|U_i \cap V_j|}{|U_i||V_j|}\right) \tag{2.15}$$

  The normalized mutual information is defined as

  $$NMI(U,V) = \frac{\mathrm{MI}(U,V)}{\sqrt{H(U)H(V)}} \tag{2.16}$$

  This value of the mutual information and also the normalized variant is not adjusted for chance and will tend to increase as the number of different labels (clusters) increases, regardless of the actual amount of mutual information between the label assignments.
  Advantages:
  - random (uniform) label assignments have a AMI score close to 0.0 for any value of $n_{clusters}$ and $n_{samples}$ (which is not the case for raw Mutual Information or the V-measure for instance)
  - bounded range $[0,1]$: Values close to zero indicate two label assignments that are largely independent, while values close to one indicate significant agreement.
    Further, values of exactly 0 indicate purely independent label assignments and a AMI of exactly 1 indicates that the two label assignments are equal
  - no assumption is made on the cluster structure: can be used to compare clustering algorithms such as k-means
- Given the knowledge of the ground truth class assignments of the samples, it is possible to define some intuitive metric using conditional entropy analysis:

- homogeneity: each cluster contains only members of a single class
- completeness: all members of a given class are assigned to the same cluster

Homogeneity and completeness scores are formally given by:

$$h = 1 - \frac{H(C|K)}{H(C)} \tag{2.17}$$

$$c = 1 - \frac{H(K|C)}{H(K)} \tag{2.18}$$

where $H(C|K)$ is the conditional entropy of the classes given the cluster assignments and is given by:

$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|} \frac{n_{c,k}}{n} \cdot \log\left(\frac{n_{c,k}}{n_k}\right) \tag{2.19}$$

and $H(C)$ is the entropy of the classes and is given by:

$$H(C) = -\sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log\left(\frac{n_c}{n}\right) \tag{2.20}$$

with n the total number of samples, n_c and n_k the number of samples respectively belonging to class c and cluster k, and finally $n_{c,k}$ the number of samples from class c assigned to cluster k.

The conditional entropy of clusters given class $H(K|C)$ and the entropy of clusters $H(K)$ are defined in a symmetric manner.

Rosenberg and Hirschberg further define V-measure as the harmonic mean of homogeneity and completeness:

$$v = 2 \cdot \frac{h \cdot c}{h + c} \tag{2.21}$$

Advantages:

- bounded scores: 0.0 is as bad as it can be, 1.0 is a perfect score
- intuitive interpretation: clustering with bad V-measure can be qualitatively analyzed in terms of homogeneity and completeness to better feel what k̈indöf mistakes is done by the assignment
- no assumption is made on the cluster structure: can be used to compare clustering algorithms such as k-means

- The Fowlkes-Mallows index can be used when the ground truth class assignments of the samples is known.

  The Fowlkes-Mallows score FMI is defined as the geometric mean of the pairwise precision and recall:

$$FMI = \frac{\text{TP}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})}} \tag{2.22}$$

Where TP is the number of True Positive (i.e. the number of pair of points that belong to the same clusters in both the true labels and the predicted

labels), FP is the number of False Positive (i.e. the number of pair of points that belong to the same clusters in the true labels and not in the predicted labels) and FN is the number of False Negative (i.e the number of pair of points that belongs in the same clusters in the predicted labels and not in the true labels).
Advantages:
  - random (uniform) label assignments have a FMI score close to 0.0 for any value of $n_{clusters}$ and $n_{samples}$ (which is not the case for raw Mutual Information or the V-measure for instance)
  - bounded range $[0, 1]$: Values close to zero indicate two label assignments that are largely independent, while values close to one indicate significant agreement. Further, values of exactly 0 indicate purely independent label assignments and a AMI of exactly 1 indicates that the two label assignments are equal (with or without permutation)
  - no assumption is made on the cluster structure: can be used to compare clustering algorithms such as k-means

- Silhouette Coefficient: measures how close a fiber is to its own cluster in comparison to the rest of the clusters, i.e. whether there is another cluster that might represent it better or as well. Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually.
  This measure has a range of [-1, 1]. Silhouette coefficients (as these values are referred to as) near $+1$ indicate that the sample is far away from the neighboring clusters.
  A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster. If the ground truth labels are not known, evaluation must be performed using the model itself.
  The Silhouette Coefficient is defined for each sample and is composed of two scores:
  - a: The mean distance between a sample and all other points in the same class.
  - b: The mean distance between a sample and all other points in the next nearest cluster.

  The Silhouette Coefficient s for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)} \tag{2.23}$$

  Advantages:
  - the score is bounded between -1 for incorrect clustering and $+1$ for highly dense clustering. Scores around zero indicate overlapping clusters
  - the score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster

- If the ground truth labels are not known, the Calinski-Harabaz index can be used to evaluate the model, where a higher Calinski-Harabaz score relates to a model with better defined clusters. For k clusters, the Calinski-Harabaz score s is given as the ratio of the between-clusters dispersion mean and the within-cluster dispersion:

$$s(k) = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N-k}{k-1} \tag{2.24}$$

  where B_K is the between group dispersion matrix and W_K is the within-cluster dispersion matrix defined by:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \tag{2.25}$$

$$B_k = \sum_{q} n_q (c_q - c)(c_q - c)^T \tag{2.26}$$

  with N be the number of points in our data, C_q be the set of points in cluster q, c_q be the center of cluster q, c be the center of E, n_q be the number of points in cluster q.
  Advantages:
  - the score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster
  - the score is fast to compute

2.4 Conclusions

The study carried out in this project aims to perform a pre-processing function of images and a support function to a clinical context. To pursue these goals, it is necessary to study how fibers are located within the brain and to study their shapes.

Understanding the geometric organization of fibers that make up the bundles within the brain is a well-studied task in neuroscience. Over the years, many methods for calculating the distance between fibers have been proposed, suitably inserted into clustering algorithms. Through our research, we want to compare all these different possibilities both at the results and at the computational levels. The objectives of this work are referred to the two main contexts of the application.

The first context is the preoperative phase, that is to provide the neurosurgeon with a simpler visualization of the fibers defining the anatomical bundles within the brain. This simplification is very important because the tractography algorithms produce a dataset consisting of a large number of fibers. Facilitating the phase of analysis for the neurosurgeon can lead to a benefit in the operative phase.

The operative phase is part of the second context of application; here it is important to assist the surgeon in tumour resection. Analyzing how the anatomical bundles are clustered and understanding which fibers are similar to each other, allow surgeons to identify by how many fibers have been deformed the presence of a tumour mass. The main contribution of this thesis is to provide neurosurgeons with new analytical tools that, together with the physician anatomical background, can provide a good framework for a new and effective pre-processing image and guideline method for the intervention. With respect to the state of the art, we have already highlighted the limits of the known similarity metrics. With our metric we show that considering together the concept of distance and shape to classify the fibers leads to improvements in the cluster phase. These improvements are directly visible in the pre-operative phase, as a metric providing a good cluster gives the neurosurgeon the opportunity to perform a more accurate analysis of the intervention. Even though we know that technology does not replace the surgeon's knowledge, we think that a good system of analysis to understand how to plan an intervention or to figure out whether an instrument is at an operational stage can have a positive impact on the patient. A positive impact on the safeguard of key functions such as word or movement, which are at risk during neurosurgery because often the tumor mass is masked by the fibers that make up the beam. Studying well the shape of the fibers and the features that cluster them brings a lower stress to the surgeon because he knows a priori the most important parts on which to concentrate.

For tractography results we analyze two types of datasets, online dataset and clinical dataset. For both, we apply some of metrics of (di)similarity know in literature and some cluster algorithms. For clustering metric validation, there is one set of standard criteria, given the knowledge of the ground truth; this criteria are indices that measure the level of performance of algorithms for brain fibers classification.

3

Mathematical Background

We present, below, the key mathematical arguments studied as a basis for our new similarity metric.

In section 3.1 we give a brief introduction explaining the metric spaces and the proprieties needed to talk about distance and the proprieties that must be satisfied for define a new metric of distance. Furthermore, the importance of similarity measures in the context of fibers clustering will be explained

In section 3.2 we introduce the concept of differential geometry of curve, for planar case and for three dimensional case. For the differential geometry of curves in the plane an essential tool is the complex structure of $\mathbb{R}^2$; we introduce a new concept of similarity between fibers as a complex number, where the Real component composed by Euclidean distance and Imaginary component, composed by angular difference. Considering the fibers of the brain as a set of points that provide generic curves, we see the limits of study in the case of fiber analysis in two dimensional space and then we analyze the behavior of the curves in 3D space; it is the space in which we make our analysis.

In section 3.3 we talk about the concept of correlation and its proprieties. Specifically we will talk about cross-correlation, the function used to classify fibers with our method.

3.1 Metric Space

A metric space [14], [62] is a set X equipped with a function d of two variables which measures the distance between points: $d(x, y)$ is the distance between two points x and y in X.

3.1.1 Metrics

A metrics on a set is a function that satisfies the minimal properties we might expect of a distance.

Definition 3.1. *A metric d on a set X is a function*

$$d : X \times X \to \mathbb{R} \qquad \textit{such that for all x,y} \in X\textit{:}$$

$$d(x,y) \geq 0 \qquad (\textit{non negativity}) \tag{3.1}$$

$$d(x,y) = 0 \quad if \quad and \quad only \quad if \quad x = y \qquad (\textit{identity of indiscernibles}) \tag{3.2}$$

$$d(x,y) = d(y,x) \qquad (\textit{symmetry}); \tag{3.3}$$

$$d(x,y) \leq d(x,z) + d(z,x) \qquad (\textit{triangle inequality}). \tag{3.4}$$

A metric space (X, d) is a set X with a metrics d defined on X.

The first condition follows from the other three. Since for any $x, y \in X$:

$$d(x,y) + d(y,x) \geq d(x,x) \qquad \text{by triangle inequality} \tag{3.5}$$

$$d(x,y) + d(x,y) \geq d(x,x) \qquad \text{by simmetry} \tag{3.6}$$

$$2d(x,y) \geq 0 \qquad \text{by identiy of indiscernibles} \tag{3.7}$$

$$d(x,y) \geq 0 \qquad \text{we have non-negativity} \tag{3.8}$$

where, function d is also called distance function.

- 1. Define: $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by

$$d(x,y) = |x - y| \tag{3.9}$$

Then d is a metric on $\mathbb{R}$. Nearly all the concepts we discuss for metric spaces are natural generalizations of the corresponding concepts for $\mathbb{R}$ with this absolute-value metrics.

- 2. Define $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ by

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \qquad x = (x_1, x_2), \quad y = (y_1, y_2) \tag{3.10}$$

Then d is a metrics on $\mathbb{R}^2$, called the Euclidean. It corresponds to the usual notion of distance between points in the plane.

- 3. The Euclidean metric $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ on $\mathbb{R}^n$ is defined by

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ...(x_n - y_n)^2} \tag{3.11}$$

where

$$x = (x_1, x_2, ..., x_n) \quad y = (y_1, y_2, ..., y_n) \tag{3.12}$$

For $n = 1$ this metrics reduces to the absolute-value metrics on $\mathbb{R}$, and for $n = 2$ it is the previous example.

- 4. Define $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ by

$$d(x, y) = max(|x_1 - y_1| + |x_2 - y_2|) \qquad x = (x_1, x_2), \quad y = (y_1, y_2) \tag{3.13}$$

Then d is a metric on $\mathbb{R}^2$, called maximum metric.

- 5. Let C(K) denote the set of continuous functions $f : K \to \mathbb{R}$, where $K \subset \mathbb{R}$ is compact; for example, we could take $K = [a, b]$ to be a closed, bounded interval. For $f, g \in C(K)$ define

$$d(f, g) = sup_{x \in K} |f(x) - g(x)| \tag{3.14}$$

The function $d : C(K) \times C(K) \to \mathbb{R}$ is well-defined, since a continuous function on a compact set is bounded; in fact, such a function attains it maximum value, so we could also write

$$d(f, g) = max_{x \in K} |f(x) - g(x)| \tag{3.15}$$

Then d is a metrics on $C(K)$. Two functions are close with respect to this metrics if their values are close at every point of K.

Subspaces of a metric space (X, d) are subsets $A \subset X$ with the metrics d_A obtained by restricting the metrics d on X to$A$.

Definition 3.2. *Let (X, d) be a metric space. A sub space (A, d_A) of (X, d) consists of a subset $A \subset X$ whose metrics $dA : A \times A \to \mathbb{R}$ is is the restriction of d to A; that is, $d_A(x, y) = d(x, y)$ for all $x, y \in A$.*

In general, there are no algebraic operations defined on a metric space, only a distance function. Most of the spaces that arise in analysis are vector, or linear, spaces, and the metrics on them are usually derived from a norm, which gives the length of a vector.

Definition 3.3. *A normed vector space $(X, || \cdot ||)$ is a vector space X, assumed to be real, together with a function $|| \cdot || : X \to \mathbb{R}$, called a norm on X, such that for all $x, y \in X$ and $k \in \mathbb{R}$:*

$$0 \leq ||x|| < \inf \quad \textit{and} \quad ||x|| = 0 \quad \textit{if and only if} \quad x = 0 \tag{3.16}$$

$$||kx|| = |k|||x|| \tag{3.17}$$

$$||x + y|| \leq ||x|| + ||y|| \tag{3.18}$$

Proposition 3.4. *If $(X, ||\cdot||)$ is a normed vector space X, then $d : X \times X \to \mathbb{R}$ defined by $d(x, y) = ||x - y||$ is a metrics on X.*

A metrics associated to a norm has the additional properties that for all $x, y, z \in X$ and $k \in \mathbb{R}$

$$d(x + z, y + z) = d(x, y), \qquad d(kx, ky) = |k|d(x, y), \tag{3.19}$$

which are called translation invariance and homogeneity, respectively. These properties do not even make sense in a general metric space since we cannot add points or multiply them by scalars.

If X is a normed vector space, we always use the metrics associated to its norm, unless stated specifically otherwise.

- The set of real numbers $\mathbb{R}$ with the absolute-value norm $|\cdot|$ is a one-dimensional normed vector space.
- The set $\mathbb{R}^2$ with any of the norms defined for $x = (x_1, x_2)$ by

$$||x||_1 = |x_1| + |x_2|, \qquad ||x||_2 = \sqrt{x_1^2 + x_2^2}, \qquad ||x||_{\inf} = max(|x_1|, (|x_2|) \tag{3.20}$$

  is a two-dimensional normed vector space.
- The set $\mathbb{R}^n$ with the $l^p - norm$ defined for $x = (x_1, x_2, ..., x_n)$ and $1 \leq p < \inf$ by

$$||x||_p = (|x_1|^p + |x_2|^p + ... + |x_n|^p)^{1/p} \tag{3.21}$$

  and for $p = \inf$ by

$$||x||_{\inf} = max(|x_1| + |x_2| + ... + |x_n|^p) \tag{3.22}$$

  is an $n-dimensional$ normed vector space for every $1 \leq p \leq \inf$. The Euclidean case $p = 2$ is distinguished by the fact that the norm $||\cdot||^2$ is derived from an inner product on $\mathbb{R}^n$:

$$||x||_2 = \sqrt{\langle x, x\rangle}, \quad \langle x, y\rangle = \sum ni = 1 x_i y_i \tag{3.23}$$

The triangle inequality for the $l^p - norm$ is called Minkowski's inequality.

- Let $K \subset \mathbb{R}$ be compact. Then the space $C(K)$ of continuous functions $f : K \to \mathbb{R}$ with the sup-norm $|| \cdot || : C(K) \to \mathbb{R}$, defined by

$$||f|| = sup_{x \in K} |f(x)| \tag{3.24}$$

is a normed vector space.

A sequence x_n in a set X is a function $f : \mathbb{N} \to X$, where we write $x_n = f(n)$ for the nth term in the sequence.

Definition 3.5. *Let (X, d) be a metric space. A sequence (x_n) in X converges to $x \in X$, written $x_n \to x$ as $n \to \inf or$*

$$\lim_{n \to \inf} x_n = x, \tag{3.25}$$

if for every $\epsilon > 0$ there exists $N \subset \mathbb{N}$ such that

$$n > N \quad implies\ that \quad d(x_n, x) < \epsilon \tag{3.26}$$

That is, $x_n \to x$ if $d(x_n, x) \to 0$ as $n \to \inf$. Equivalently, $x_n \to x$ as $n \to \inf$ if for every neighborhood U of x there exists $N \in N$ such that $x_n \in U$ for all $n > N$.

- Let $K \subset \mathbb{R}$ be compact. A sequence of continuous functions (f_n) in $C(K)$ converges to $f \in C(K)$ with respect to the sup-norm if and only if $f_n \to f$ as $n \to \inf$ uniformly on K.

We define closed sets in terms of sequences in the same way as for $\mathbb{R}$.

Definition 3.6. *A subset $F \subset X$ of a metric space X is sequentially closed if the limit every convergent sequence (x_n) in F belongs to F.*

Explicitly, this means that if (x_n) is a sequence of points $x_n \in F$ and $x_n \to x$ as $n \to \inf$ in X, then $x \in F$. A subset of a metric space is sequentially closed if and only if it is closed.

- Let $F \subset C(K)$ be the set of continuous functions $f : K \to \mathbb{R}$ such that $|f(x)| \leq 1$ for all $x \in K$. Then F is a closed subset of $C(K)$.

We can also give a sequential definition of compactness, which generalizes the Bolzano-Weierstrass property.

Definition 3.7. *A subset $K \subset X$ of a metric space X is sequentially compact if every sequence in K has a convergent subsequence whose limit belongs to K.*

Explicitly, this means that if (x_n) is a sequence of points $x_n \in K$ then there is a subsequence (x_{n_k}) such that $x_{n_k} \to x$ as $k \to \inf$, and $x \in K$.

Theorem 3.8. *A subset of a metric space is sequentially compact if and only if it is compact.*

We can also define Cauchy sequences in a metric space.

Definition 3.9. *Let (X, d) be a metric space. A sequence (x_n) in X is a Cauchy sequence for every $\epsilon > 0$ there exists $N \in N$ such that $m, n > N$ implies that $d(x_m, x_n) < \epsilon$.*

Completeness of a metric space is defined using the Cauchy condition.

Definition 3.10. *A metric space is complete if every Cauchy sequence converges.*

For $\mathbb{R}$, completeness is equivalent to the existence of supreme, but general metric spaces are not ordered so this property does not apply to them.

3.1.2 Similarity Metric and Clustering problem

From the scientific and mathematical point of view, distance is defined as a quantitative degree of how far apart two objects are. Synonyms for distance include dissimilarity. Those distance measures satisfying the metric properties are simply called metrics while other non-metric distance measures are occasionally called divergence [13]. Synonyms for similarity include proximity, and similarity measures are often called similarity coefficients.

The similarity measure is the measure of how much alike two data objects are. Similarity measure in a data mining context is a distance with dimensions representing features of the objects. If this distance is small, it will be the high degree of similarity where large distance will be the low degree of similarity [73]. The similarity is subjective and is highly dependent on the domain and application. For example, two fruits are similar because of color or size or taste. Care should be taken when calculating distance across dimensions/features that are unrelated. The relative values of each element must be normalized, or one feature could end up dominating the distance calculation. Similarity are measured in the range 0 to 1. Similarity is 1 if $X = Y$, is 0 otherwise.

The input to a clustering problem is a finite nonempty set P together with finitely many attributes, or a dissimilarity coefficient (DC). The goal of a clustering algorithm is to produce some sort of nested sequence of classifications of P. Most (though not all) algorithms proceed by first constructing a DC, and then using the DC to produce the classifications. Our efforts will largely lie with that part of the procedure that goes from a DC to the output classifications. First we need some definitions. We agree to let $[0, \infty)$ denote the nonnegative real numbers, and $\sum : (P)$ the set of reflexive, symmetric relations on P, ordered by inclusion.

Definition 3.11. *A dissimilarity coefficient (DC) is a mapping $d : P \times P \to [0, \infty)$ that satisfies:*

$$d(x,x) = 0 \quad \textit{for all } x \quad \in P \tag{3.27}$$

$$d(x,y) = d(y,x) \quad \textit{for all } x,y \quad \in P \tag{3.28}$$

it is called an ultrametric if it also satisfies:

$$d(x,y) \leq max\{d(x,z), d(y,z)\} \quad \textit{for all } x,y,z \quad \in P \tag{3.29}$$

Definition 3.12. *A numerically stratified clustering* (NSC) *is a mapping* $C : [0,\infty) \mapsto \sum(P)$

that satisfies:

$$h \leq k \quad \textit{implies} \quad C(h) \subseteq C(k) \tag{3.30}$$

$$\textit{there exists} \quad h \geq 0 \quad \textit{such that} \quad C(h) = P \times P \tag{3.31}$$

corresponding to any $h \geq 0$,

$$\textit{there is a number} \quad \delta = \delta(h) > 0 \quad \textit{such that} \quad C(h) = C(h+\delta) \quad \textit{(continuity from the right)} \tag{3.32}$$

$$if \quad C(h) \quad \textit{is an equivalence relation for every} \quad h \in [0,\infty), thenC \quad \textit{is called a dendrogram} \tag{3.33}$$

It will be convenient to let $D(P)$ denote the set of dissimilarity coefficients on P, and $NSC(P)$ the set of numerically stratified clusterings. $D(P)$ is ordered by the rule $d1 \leq d2$ if and only if $d1(x,y) \leq d2(x,y)$ for all $x, y \in P$, and $NSC(P)$ is ordered in the analogous manner.

A cluster algorithm can then be thought of as a transformation from $D(P)$ into $NSC(P)$, where the output is usually a dendrogram.

Theorem 3.13. *(i) corresponding to each DC d there is an NSC* Td *given by*

$$Td(h) = \{(x,y) : d(x,y) \leq h\} \quad (h \geq 0) \tag{3.34}$$

(ii) corresponding to each NSC C, there is a DC d_C *given by*

$$d_C(x,y) = min\{h \in [0,\infty) : (x,y) \in C(h)\} \quad (x, y \in P) \tag{3.35}$$

(iii) the correspondence $d \mapsto Td$ *is a bijection between* $D(P)$ *and* $NSC(P)$ *whose inverse is the mapping* $C \mapsto d_C$

(iv) the DC d is an ultrametric if and only if Td *is a dendrogram*

3.2 Curves

The differential geometry of curves and surfaces has two aspects. One, which may be called classical differential geometry [68], [70], [65], started with the beginnings of calculus. The classical differential geometry analyzes the local properties of the curves and surfaces by using methods based on differential calculus. Thus, curves and surfaces are defined by functions that can be differentiated a certain number of times. As classical differential geometry represents mostly the study of surfaces, the local properties of curves are an important part of this, since they appear naturally while studying surfaces.

The other aspect is the so-called global differential geometry [11]. The global differential geometry on the other hand, studies the influence of the local properties on the behavior on the entire curve or surface, on contrary to the classical geometry that studies the local properties that depend only on the behavior of the curve or surface in the neighborhood of a point. The local properties correspond to the characteristics of the object in the neighborhood of a point on the object such as the curvature of a curve at the point, while the global properties correspond to the characteristics of the object on a large scale and over extended parts of the object such as the number of stationary points of a curve.

Another classification of the properties of curves, which is based on their relation to the embedding external space which they reside in, may be made where the properties are divided into intrinsic and extrinsic.

The first category corresponds to those properties which are independent in their existence and definition from the ambient space which embraces the object such as the distance along a given curve or the Gaussian curvature; the second category is related to those properties which depend in their existence and definition on the external embedding space such as having a normal vector at a point on the curve.

The intrinsic properties are defined and expressed in terms of the metric tensor which is formulated in differential geometry as the first fundamental form while the extrinsic properties are expressed in terms of the surface curvature tensor which is formulated in differential geometry as the second fundamental form.

3.2.1 Curves in the plane.

Properties of Euclidean space.

Definition 3.14. *Euclidean n-space $\mathbb{R}^n$ consists of the set of all real n-tuples:*

$$\mathbb{R}^n = \{(p_1, ..., p_n) | p_j \quad \textit{is a real number for} \quad j = 1, ..., n\} \tag{3.36}$$

The elements of $\mathbb{R}^n$ are called vectors.

$\mathbb{R}^n$ is an n-dimensional vector space; this means that the operations of addition and scalar multiplication are defined. Thus if:

$$p = (p_1, ..., p_n) \quad and \quad q = (q_1, ..., q_n), \tag{3.37}$$

then $p + q$ is the element of $\mathbb{R}^n$ given by

$$p + q = (p_1 + q_1, ..., p_n + q_n). \tag{3.38}$$

Similarly, for $\lambda \in \mathbb{R}$ the vector λp is defined by

$$\lambda p = (\lambda p_1, ..., \lambda p_n). \tag{3.39}$$

The dot product of $\mathbb{R}^n$, $\cdot$, is an operation that assigns to each pair of vectors $p = (p_1, ..., p_n)$ and $q = (q_1, ..., q_n)$ the real number

$$p \cdot q = \sum_{j=1}^{n} p_j q_j. \tag{3.40}$$

The norm and distance functions of $\mathbb{R}^n$ are defined by

$$||p|| = \sqrt{p \cdot q} \tag{3.41}$$

$$distance(p, q) = ||p - q|| \tag{3.42}$$

for $p, q \in \mathbb{R}^n$. These functions have the following properties:

$$p \cdot q = q \cdot p, \qquad (p + r) \cdot q = p \cdot q + r \cdot q, \tag{3.43}$$

$$(\lambda p) \cdot q = \lambda (p \cdot q) = p \cdot (\lambda q), \qquad ||\lambda p|| = |\lambda| ||p||, \tag{3.44}$$

for $\lambda \in \mathbb{R}$ and $p, q, r \in \mathbb{R}^n$. The Cauchy-Schwarz and triangle inequalities state that for all $p, q \in \mathbb{R}^n$ we have

$$|p \cdot q| \leq ||p|| ||q|| \quad and \quad ||p + q|| \leq ||p|| + ||q||. \tag{3.45}$$

The angle θ between nonzero vectors $p, q \in \mathbb{R}^n$, is defined by the conditions:

$$cos\theta = \frac{p \cdot q}{||p||||q||}, \quad 0 \leq \theta \leq pi. \tag{3.46}$$

The Cauchy-Schwarz inequality implies that

$$-1 \leq \frac{p \cdot q}{||p||||q||} \leq 1 \tag{3.47}$$

for nonzero $p, q \in \mathbb{R}^n$.

A linear map of $\mathbb{R}^n$ into $\mathbb{R}^n$ is a function $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that

$$A(\lambda p + \mu q) = \lambda A p + \mu A q \tag{3.48}$$

for $\lambda, \mu \in \mathbb{R}$ and $p, q \in \mathbb{R}^n$.

For the differential geometry of curves in the plane an essential tool is the complex structure of $\mathbb{R}^2$; it is the linear map $J : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ given by

$$J(p_1, p_2) = (-p_2, p_1). \tag{3.49}$$

It is easy to show that the complex structure J has the following properties:

$$J^2 = -I \tag{3.50}$$

$$(Jp)(Jq) = p \cdot q \tag{3.51}$$

$$(Jp) \cdot p = 0 \tag{3.52}$$

for $p, q \in \mathbb{R}^2$, where $I : \mathbb{R}^2 \rightarrow R^2$ is the identity linear map. It is possible to use the complex structure J of $\mathbb{R}^2$ to define the signed curvature of a plane curve.

A point in the plane $\mathbb{R}^2$ can be considered a complex number via the canonical isomorphism

$$p = (p_1, p_2) \longleftrightarrow p_1 + i p_2 = \Re(p) + i\Im(p), \tag{3.53}$$

where $\Re(p)$ and $\Im(p)$ denote the real and imaginary parts of p. We shall need descriptions of the dot product $\cdot$ and the complex structure J in terms of complex numbers. Recall that the complex conjugate and absolute value of a complex number p are defined by

$$\bar{p} = \Re(p) - i\Im(p) \quad and \quad |p| = \sqrt{p\bar{p}} \tag{3.54}$$

Lemma 3.15. *Identify the plane R^2 with the set of complex numbers $\mathbb{C}$, and let $p, q \in R^2 = \mathbb{C}$. Then*

$$Jp = ip, \qquad |p| = ||q|| \quad and \quad p\bar{q} = p \cdot q + i(p \cdot Jq) \tag{3.55}$$

The angle between vectors in R^n does not distinguish between the order of the vectors, but there is a refined notion of angle between vectors in R^2 that makes this distinction.

Lemma 3.16. *Let p and q be nonzero vectors in R^2. There exists a unique θ with the following properties:*

$$cos\theta = \frac{p \cdot q}{||p||||q||}, \quad sin\theta = \frac{p \cdot Jq}{||p||||q||}, \quad 0 \leq \theta < 2\pi \tag{3.56}$$

We call θ the oriented angle from p to q.

Proof. Since $(p\bar{q})/(|p||q|)$ is a complex number of absolute value 1, it lies on the unit circle in $\mathbb{C}$; thus there exists a unique θ with $0 \leq \theta < 2\pi$ such that

$$\frac{p\bar{q}}{|p||q|} = e^{i\theta} \tag{3.57}$$

We find that

$$p \cdot q + i(p \cdot Jq) = e^{i\theta}|p||q| = |p||q|cos\theta + i|p||q|sin\theta \tag{3.58}$$

Curves in $\mathbb{R}^n$: the algebraic properties.

Definition 3.17. *Let $\alpha : (a, b) \to \mathbb{R}^n$ be a function, where (a, b) is an open interval in $\mathbb{R}$*

$$\alpha(t) = (a_1(t), ..., a_n(t)), \tag{3.59}$$

with a_j ordinary real-valued function of a real variable. The parameter α is differentiable provided a_j is differentiable $forj = 1, ..., n$.

Definition 3.18. *A parametrized curve in $\mathbb{R}^n$ is a differentiable function*

$$\alpha : (a, b) \to \mathbb{R}^n, \tag{3.60}$$

where (a, b) is an open interval in $\mathbb{R}$. If I is any other subset of $\mathbb{R}$:

$$\alpha : I \to \mathbb{R}^n \tag{3.61}$$

is a curve provided there is an open interval (a, b) containing I such that α can be extended as a differentiable function from (a,b) into $\mathbb{R}^n$.

Definition 3.19. *Let $\alpha : (a, b) \to \mathbb{R}^n$ be a curve with $\alpha(t) = (a_1(t), ..., a_n(t))$. Then the velocity of α is the function $\alpha^{'} : (a, b) \to \mathbb{R}^n$ given by:*

$$\alpha^{'}(t) = (a_1^{'}(t), ..., a_n^{'}(t)) \tag{3.62}$$

The function v defined by $v(t) = ||\alpha^{'}(t)||$ is called the speed of α. The acceleration of α is $\alpha^{''}$.

Definition 3.20. *A curve $\alpha : (a, b) \to \mathbb{R}^n$ is said to be regular if it is differentiable and its velocity is everywhere defined and nonzero. If $||\alpha^{'}(t)|| = 1$ for $a < t < b$, then α is said to have unit-speed.*

One of the most important geometric quantities associated with a curve is its length.

Definition 3.21. *Let $\alpha : (a, b) \to \mathbb{R}^n$ be a curve. Assume that α is defined on a slightly larger interval containing (a, b), so that α is defined and differentiable at a and b. Then the length of α over the interval $[a, b]$ is given by:*

$$length[a, b][\alpha] = length[\alpha] = \int_a^b ||\alpha^{'}(t)||dt \tag{3.63}$$

The length does not depend on the parametrization of the curve.

Definition 3.22. *Fix a number c with $a < c < b$. The arc length function s starting at c of a curve $\alpha : (a, b)) \to \mathbb{R}^n$ is defined by*

$$s(t) = \int_c^t ||\alpha^{'}(u)||du, \tag{3.64}$$

for $c \leq t \leq b$.

Theorem 3.23. *Let $\alpha : (a, b) \to \mathbb{R}^n$ be a regular curve. Then there exists a unit-speed reparametrization β of α.*

Proof. By the fundamental theorem of calculus, any arc length function s of α satisfies:

$$\frac{ds}{dt}(t) = s^{'}(t) = ||\alpha^{'}(t)|| \tag{3.65}$$

Since α is regular, by definition $\alpha^{'}(t)$ is never zero; hence ds/dt is always positive.
Define β by $\beta(s) = \alpha(t(s))$. By:

Lemma 3.24. *The chain rule for curves. Suppose that β is a reparametrization of α. Write $\beta = \alpha \circ h$, where $h : (c, d) \to (a, b)$ is differentiable. Then:*

$$\beta^{'}(u) = h^{'}(u)\alpha^{'}(h(u)) \tag{3.66}$$

for $c < u < d$.

then:

$$\beta^{'}(s) = (dt/ds)\alpha^{'}(t(s)) \tag{3.67}$$

Hence:

$$||\beta^{'}(s)|| = ||\frac{dt}{ds}\alpha^{'}(t(s))|| = \frac{dt}{ds}||\alpha^{'}(t(s))|| = \frac{dt}{ds}(s)\frac{ds}{dt}(t(s)) = 1 \tag{3.68}$$

The arc length function of any unit-speed curve $\beta : (c, d) \to \mathbb{R}^n$ starting at c satisfies:

$$s = s(t) = \int_c^t ||\beta^{'}(u)||du = t - c \tag{3.69}$$

Thus the function s actually measures length along β. This is the reason why unit-speed curves are said to be parametrized by arc length.

Vector Fields along Curves

Definition 3.25. *Let $\alpha : (a, b) \to \mathbb{R}^n$ be a curve. A vector field along α is a function Y that assigns to each t with $a < t < b$ a vector $Y(t)$ at the point $\alpha(t)$. At this beginning stage we shall not distinguish between a vector at $\alpha(t)$ and the vector parallel to it at the origin. This means a vector field Y along a curve α is really an n-tuple of functions:*

$$Y(t) = (y_1(t), ..., y_n(t)) \tag{3.70}$$

Thus differentiability of Y means that each of the functions $y_1, ..., y_n$ is differentiable; If Y is a differentiable vector field along a curve, then the derivative $Y^{'}$ is defined by

$$Y^{'}(t) = (y_1^{'}(t), ..., y_n^{'}(t)) \tag{3.71}$$

Lemma 3.26. *Let X and Y be differentiable vector fields along a curve $\alpha : (a, b) \rightarrow \mathbb{R}^n$, and let $f : (a, b) \rightarrow \mathbb{R}$ be differentiable. Then*

(i) $(fY)^{'} = f^{'}Y + fY^{'}$

(ii) $(X + Y)^{'} = X^{'} + Y^{'}$

(iii) $(X \cdot Y)^{'} = X^{'} \cdot Y^{'} + Y^{'} \cdot X^{'}$

(iv) $(JY)^{'} = JY^{'}$ *(for $n = 2$)*

Curvature of curves in the plane

Definition 3.27. *Let $\alpha : (a, b) \rightarrow \mathbb{R}^2$ be a regular curve. The curvature $k[\alpha]$ of α is given by the formula:*

$$k[\alpha](t) = \frac{\alpha^{''}(t) \cdot J\alpha^{'}(t)}{||\alpha^{'}(t)||^3} \tag{3.72}$$

The positive function $1/|k[\alpha]|$ is called the radius of curvature of α.

The curvature can assume both positive and negative values. In many cases k is called the signed curvature Figure(3.1) in order to distinguish it from the curvature in 3D space.

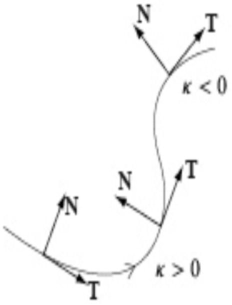


Fig. 3.1: Curvature
Definition of curvature with signed.

Like the function *length*$[\alpha]$ is a geometric quantity associated with a curve α and it is independent of the parametrization.

Theorem 3.28. *Let $\alpha : (a,b) \to \mathbb{R}^2$ be a regular curve, and let $\beta : (c,d) \to \mathbb{R}^2$ be a reparametrization of α. Write $\beta = \alpha \circ h$, where $h : (c,d) \to (a,b)$ is differentiable. Then:*

$$k[\beta](t) = (sign(h^{'}(t)))k[\alpha](h(t)), \tag{3.73}$$

wherever $sign(h^{'}(t))$ is defined. The curvature of a plane curve is up to sign independent of the parametrization.

Theorem 3.29. *Let $\alpha : (a,b) \to \mathbb{R}^2$ be a regular curve.*

i) α is part of a straight line if and only if $k[\alpha](t) \equiv 0$.

ii) α is part of a circle of radius $a > 0$ if and only if $|k[\alpha]|(t) \equiv 1/a$.

3.2.2 Curves in Space

The Vector Cross Product on $\mathbb{R}^3$

Let us recall the notion of vector cross product on $\mathbb{R}^3$.First, let us agree to use the notiation

$$i = (1,0,0), \qquad j = (0,1,0), \qquad k = (0,0,1) \tag{3.74}$$

For $a \in \mathbb{R}^3$ we can write either $a = (a_1, a_2, a_3)$ or $a = a_1 i + a_2 j + a_3 k$.

Definition 3.30. *Let $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ be vectors in $\mathbb{R}^3$. Then the vector cross product of a and b is given by*

$$a \times b = det \begin{cases} i & j \quad k \\ a_1 & a_2 \quad a_3 \\ b_1 & b_2 \quad b_3 \end{cases}$$

The vector cross product on $\mathbb{R}^3$ has the following well-known properties:

(i) $\quad a \times c = -c \times a$

(ii) $\quad (a+b) \times c = a \times c + b \times c$

(iii) $\quad (\lambda a) \times c = \lambda(a \times c) = a \times (\lambda c)$

(iv) $\quad (a \times b) \cdot (c \times d) = (a \cdot c)(b \cdot d) - (a \cdot d)(b \cdot c)$

$(v) \quad ||a \times b||^2 = ||a||^2||b||^2 - (a \cdot b)^2$

$(vi) \quad a \times (b \times c) = (a \cdot c)b - (a \cdot b)c$

$(vii) \quad (a \times b) \cdot c = a \cdot (b \times c)$

for $\lambda \in \mathbb{R}$ and $a, b, c, d \in \mathbb{R}^3$. For $a, b, c \in \mathbb{R}^3$ we define the vector triple product (abc) by

$$(abc) = det \begin{cases} a \\ b \\ c \end{cases} = \text{det} \begin{cases} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{cases}$$

where $a = (a_1, a_2, a_3)$ and so forth. The vector triple product is related to the dot product and the cross product by the formulas

$$(abc) = (a \times b) \cdot c = a \cdot (b \times c) \tag{3.75}$$

Finally, we note that

$$||a \times b|| = sin\theta||a||||b|| \tag{3.76}$$

where θ is the positive angle between a and b.

Curvature and Torsion of Curves in $\mathbb{R}^3$

The notion of curvature of a unit-speed curve in $\mathbb{R}^n$ for general n is defined by:

Definition 3.31. *Let $\beta : (c, d) \to \mathbb{R}^n$ be a unit-speed curve. Write*

$$k[\beta](s) = ||\beta^{''}(s)|| \tag{3.77}$$

Then the function $k[\beta] : (c, d) rightarrow \mathbb{R}$ is called the curvature of β.

Curvature measures the failure of a curve to be a straight line. A straight line in $\mathbb{R}^n$ is characterized by the fact that its curvature vanishes.

Lemma 3.32. *Let $\beta : (c, d) \to \mathbb{R}^n$ be a unit-speed curve. The following conditions are equivalent:*

$(i) \quad k \equiv 0$

$(ii) \quad \beta'' \equiv 0$

(iii) *β is a straight line.*

Definition 3.33. *Let $\beta : (c, d) \to \mathbb{R}^n$ be a unit-speed curve. Then*

$$T = \beta^{'} \tag{3.78}$$

is called the unit tangent vector field of β.

Lemma 3.34. *If F is a vector field of unit length along a curve $\alpha : (a, b) \to \mathbb{R}^n$, then $F^{'} \cdot F = 0$.*

Proof.

$$F(t)\dot{F}(t) = 1 \tag{3.79}$$

for all t. Differentiating, $2F(t) \cdot F^{'}(t) = 0$. cvd

Taking $F = T$, for Lemma 1.13, $T^{'} \cdot T = 0$.

In the unit-speed curves in $\mathbb{R}^3$ the curvature is strictly positive. This implies that $T^{'} = \beta^{''}$ never vanishes.

Definition 3.35. *Let $\beta : (c, d) \to \mathbb{R}^3$ be a unit-speed curve with $k(s) > 0$ for $c < s < d$. The vector field:*

$$N = \frac{T^{'}}{k} \tag{3.80}$$

is called the principal normal vector field and $B = T \times N$ is called the binormal vector field. The triple T, N, B is called the Frenet frame field on β.

This frame forms one of the basic tools for the differential geometry of space curves.

Theorem 3.36. *Let $\beta : (c, d) \to \mathbb{R}^3$ be a unit-speed curve with $k(s) > 0$ for $c < s < d$. Then:*

i) $||T|| = ||N|| = ||B|| = 1$ and $T \cdot N = N \cdot B = B \cdot T = 0$.

ii) Any vector field F along β can be expanded as:

$$F = (F \cdot T)T + (F \cdot N)N + (F \cdot B)B \tag{3.81}$$

iii) The Frenet Formulas hold:

$$\begin{aligned} T^{'} &= \quad kN \\ N^{'} &= -kT \quad \tau B \\ B^{'} &= \quad \tau N \end{aligned} \tag{3.82}$$

where $\tau = \tau[\beta]$ is called the torsion of the curve β.

Definition 3.37. *Let $\alpha : (a, b) \to \mathbb{R}^3$ be a regular curve, and let $\tilde{\alpha} : (c, d) \to \mathbb{R}^3$ be a unit-speed reparametrization of α. Write $\alpha(t) = \tilde{\alpha}(s(t))$, where $s(t)$ is the arc length function. Denote by $\tilde{k}$and $\tilde{\tau}$ the curvature and torsion of $\tilde{\alpha}$, respectively. Also, let $\{\tilde{T}, \tilde{N}, \tilde{B}, \}$ be the Frenet frame field of $\tilde{\alpha}$.*

Then we define:

$$k(t) = \tilde{k}(s(t)), \qquad \tau(t) = \tilde{\tau}(s(t), \tag{3.83}$$

$$T(t) = \tilde{T}(s(t)), \qquad N(t) = \tilde{N}(s(t)), \qquad B(t) = \tilde{B}(s(t)) \tag{3.84}$$

In other words, the curvature, torsion and Frenet frame field of an arbitrary-speed curve α are reparametrizations of those of a unit-speed parametrization of α.

Theorem 3.38. *Let $\alpha : (a, b) \to \mathbb{R}^3$ be a regular curve with speed $v = ||\alpha^{'}|| = s^{'}$. Then the following generalizations of the Frenet formulas hold:*

$$\begin{aligned} T^{'} &= \qquad vkN \\ N^{'} &= -vkT \quad + v\tau B \\ B^{'} &= \qquad v\tau N \end{aligned} \tag{3.85}$$

Proof. By the chain rule we have:

$$\begin{aligned} T^{'}(t) = s^{'}(t)\tilde{T}^{'}(s(t)) = v(t)\tilde{T}^{'}(s(t)) \\ N^{'}(t) = s^{'}(t)\tilde{N}^{'}(s(t)) = v(t)\tilde{N}^{'}(s(t)) \\ B^{'}(t) = s^{'}(t)\tilde{B}^{'}(s(t)) = v(t)\tilde{B}^{'}(s(t)) \end{aligned} \tag{3.86}$$

Thus from the Frenet formulas for $\tilde{\alpha}$, it follows that:

$$T^{'} = v(t)\tilde{k}(s(t))\tilde{N}(s(t)) = v(t)k(t)N(t) \tag{3.87}$$

The other two formulas are proved similarly.

Lemma 3.39. *The velocity $\alpha^{'}$ and acceleration $\alpha^{''}$ of a regular curve α are given by*

$$\alpha^{'} = vT, \tag{3.88}$$

$$\alpha^{''} = \frac{dv}{dt}T + v^2kN, \tag{3.89}$$

where v denotes the speed of α.

Proof. Write $\alpha(t) = \tilde{\alpha}(s(t))$, where $\tilde{\alpha}$ is a unit-speed parametrization of α. By the chain rule we have:

$$\alpha(t) = \tilde{\alpha}(s(t))s^{'}(t) = v(t)\tilde{T}(s(t)) = v(t)T(t), \tag{3.90}$$

$$\alpha'' = \frac{dv}{dt}T + vT^{'} = \frac{dv}{dt}T + v^2kN \tag{3.91}$$

Theorem 3.40. *Let $\alpha : (a,b) \to \mathbb{R}^3$ be a regular curve with nonzero curvature. Then:*

$$\begin{aligned} T =& \frac{\alpha^{'}}{||\alpha^{'}}||, \\ N =& B \times T, \\ B =& \frac{\alpha^{'} \times \alpha^{''}}{||\alpha' \times \alpha''||}, \\ k =& \frac{||\alpha^{'} \times \alpha^{''}||}{||\alpha^3||}, \\ \tau =& \frac{\alpha^{'} \times \alpha^{''} \cdot \alpha^{'''}}{||\alpha' \times \alpha''||^2} \end{aligned} \tag{3.92}$$

We have attached three orthonormal vectors T (s), N (s), B(s) to each point of the curve [52], [63], [12]. These three vectors form a basis of R^3, which moves along the curve. Geometers use the word frame instead of the word basis, so T,N,B is called the moving frame. Imagine an isolated point traveling along the curve at constant speed. The moving frame allows us to replace this single point with an airplane. The nose of the airplane should point along T, the left wing along N, and the tail along B. As we travel along the curve, the airplane pitches and rolls as T,N, and B move. At each point, the orientation of the plane is completely determined by T,N, and B.

Theorem 3.41. *(The Fundamental Theorem of Space Curves) Let α and γ be unit-speed curves in $\mathbb{R}^3$ defined on the same interval (a,b), and assume they have the same torsion and the same positive curvature. Then there is a Euclidean motion F of $\mathbb{R}^3$ that maps α onto γ.*

3.3 Correlation

The examples are vectors in $\mathbb{R}^n$ but they represent an arbitrary shift of periodic time series.

The following notation: y_{+s} refers to the vector y shifted by s positions, where positions are shifted modulo n. We then use the standard inner product between shifted examples:

$$\langle x, y_{+s} \rangle = \sum_{i=1}^{n} x_i (y_{+s})_i \tag{3.93}$$

In the context of time series, computing the cross-correlation corresponds to aligning two time series such that their inner product, or similarity, is maximized.

3.3.1 Properties of Cross-Correlation

Cross-Correlation [36], [2] has properties similar to an inner product, and can be used intuitively as a similarity function:

Theorem 3.42. *Cross-Correlation*

(i) $C(x,x) = \langle x, x \rangle \geq 0$

(ii) $C(x,y) = C(y,x)$

(iii) $C(x,y) \leq \sqrt{C(x,x)C(y,y)}$ *per ogni* x, y *The Cauchy-Schwartz Inequality*

(iv) If we use the cross-correlation function to give a distance measure d *such that:*

$$d(x,y)^2 = C(x,x) + C(y,y) - 2C(x,y) = min_s ||x - (y_{+s})||^2$$

then d *satisfies the Triangle Inequality.*

Proof. For (i) note that by definition $C(x,x) \geq \langle x, x \rangle$. On the other hand, $C(x,x) = \sum x_i x_{i+s}$ and by the Cauchy-Schwartz inequality:

$$\sum x_i x_{i+s} \leq \sqrt{\sum x_i^2} \sqrt{\sum x_{i+s}^2} = \sqrt{\sum x_i^2} \sqrt{\sum x_i^2} = \langle x, x \rangle$$

which means $\langle x, x \rangle \geq C(x,x) \geq \langle x, x \rangle$ or $C(x,x) = \langle x, x \rangle \geq 0$

To prove (ii) observe that since $\langle x, y_{+s} \rangle = \langle x_{-s}, y \rangle = \langle x_{+(n-s)}, y \rangle$ maximizing over the shift for y is the same as maximizing over the shift for x.

Let $C(x,y) = \langle x, y_{+s} \rangle = \langle x, z \rangle$, where s is the shift maximizing the correlation and where we denote $z = y_{+s}$.

Then by, (i), $\sqrt{C(x,x)C(y,y)} = \sqrt{\langle x, x \rangle \langle y, y \rangle} = ||x|| ||y||$.

Therefore the claim is equivalent to $||x|| ||y|| \geq \langle x, z \rangle$ and since the norm does not change under shifting the claim is equivalent to $||x|| ||z|| \geq \langle x, z \rangle = C(x,y)$.

The last inequality holds by the Caucht-Schwartz inequality for normal inner products.

For (iv), let $x, y, z \in \mathbb{R}^n$. Let τ_{ab} be the shift that minimizes $d(a, b)$.

$$\begin{aligned}
d(x,y) + d(y,z) &= ||(x_{+\tau_{xy}}) - y|| + ||(y_{+\tau_{yz}}) - z|| \quad (1)\\
&= ||(x_{+\tau_{xy}+\tau_{xy}}) - (y_{+\tau_{yz}})|| + ||(y_{+\tau_{yz}}) - z|| \quad (2)\\
&\geq ||(x_{+\tau_{xy}+\tau_{xy}}) - (y_{+\tau_{yz}}) + (y_{+\tau_{yz}}) - z|| \quad (3)\\
&= ||(x_{+\tau_{xy}+\tau_{xy}}) - z|| \quad (4)\\
&\geq ||(x_{+\tau_{xy}}) - z|| = d(x,z) \quad (5)
\end{aligned}$$

Where (2) holds because shifting x and y by the same amount does not change the value of $||x - y||$, (3) holds because of the triangle inequality, and (5) holds because by definition τ_{xz} minimizes the distance between x and z.

3.3.2 Autocorrelation

The autocorrelation of $X(t)$, denoted by $r_{XX}(t_1, t_2)$, is defined as the expected value of the product of the values of the process at two different time points as follows:

$$r_{XX}(t_1, t_2) = E\{X(t_1)X^*(t_2)\} \tag{3.94}$$

The complex conjugate in the above definition assures that the product becomes the square of the magnitude of the second moment of $X(t)$ when the two time points coincide, $t_1 = t_2$.

By substituting $X(t_1)$and $X^*(t_2)$ into the above definition and expanding it, we obtain the following expression:

$$\begin{aligned}
r_{XX}(t_1,t_2) =&E[\{X_r(t_1) + j * X_i(t_1)\}\{X_r(t_2) - j * X_i(t_2)\}]\\
=&E[\{X_r(t_1)X_r(t_2) + X_i(t_1)X_i(t_2)\} + ...\\
...+&j * \{X_i(t_1)X_r(t_2) - X_r(t_1)X_i(t_2)\}]\\
=&E[\{X_r(t_1)X_r(t_2)\} + E\{X_i(t_1)X_i(t_2)\}] + ...\\
...+&j * [E\{X_i(t_1)X_r(t_2)\} - E\{X_r(t_1)X_i(t_2)\}]
\end{aligned} \tag{3.95}$$

The first two expectation operations in the above equation are the autocorrelations defined by () for the real and imaginary components of $X(t)$. The third and fourth expectation operations in the above equation are the cross-correlations between the real and imaginary components of $X(t)$, which will be defined later by (). The above equation can be written in terms of these autocorrelation and cross-correlation functions as follows:

$$\begin{aligned}
r_{XX}(t_1,t_2) =&[r_{X_rX_r}(t_1,t_2) + [r_{X_iX_i}(t_1,t_2)] + j * [r_{X_iX_r}(t_1,t_2) + [r_{X_rX_i}(t_1,t_2)]\\
=&r^r_{XX}(t_1,t_2) + j * r^i_{XX}(t_1,t_2)
\end{aligned} \tag{3.96}$$

where the superscripts r and i denote the real and imaginary components of the autocorrelation of $X(t)$:

$$\begin{aligned}
r_{XX}(t_1,t_2) =&[r_{X_rX_r}(t_1,t_2) + [r_{X_iX_i}(t_1,t_2)] + j * [r_{X_iX_r}(t_1,t_2) + [r_{X_rX_i}(t_1,t_2)]\\
=&r^r_{XX}(t_1,t_2) + j * r^i_{XX}(t_1,t_2)
\end{aligned} \tag{3.97}$$

3.3.3 Cross-correlation

The cross-correlation of $X(t)$ and $Y(t)$ is defined by:

$$r_{XY}(t_1, t_2) = E\{X(t_1)Y^*(t_2)\} \tag{3.98}$$

Expanding the above equation, we obtain the following expression:

$$\begin{aligned} r_{XY}(t_1, t_2) &= E[\{X_r(t_1) + jX_i(t_1)\}\{Y_r(t_2) - jY_i(t_2)\} \\ &= E[\{X_r(t_1)[\{Y_r(t_2) + X_i(t_1)Y_i(t_12)\} + jX_i(t_1)Y_r(t_2) - X_r(t_1)Y_i(t_12)\}] \\ &= [E\{X_r(t_1)Y_r(t_2)\} + E\{X_i(t_1)Y_i(t_2)\}] + j[E\{X_i(t_1)Y_r(t_2)\} - E\{X_r(t_1)Y_i(t_2)\}] \end{aligned} \tag{3.99}$$

The four expected values in the above equation are the cross-correlations defined by (1.99) for the real and imaginary components of $X(t)$ and the cross-correlation becomes the following equation:

$$\begin{aligned} r_{XY}(t_1, t_2) =& [r_{X_rY_r}(t_1, t_2) + r_{X_iY_i}(t_1, t_2)] + j[r_{X_iY_r}(t_1, t_2) - r_{X_rY_i}(t_1, t_2)] \\ =& r^r_{XY}(t_1, t_2) + jr^i_{XY}(t_1, t_2) \end{aligned} \tag{3.100}$$

where

$$r^r_{XY}(t_1, t_2) = r_{X_rY_r}(t_1, t_2) + r_{X_rY_r}(t_1, t_2) \tag{3.101}$$

$$r^i_{XY}(t_1, t_2) = r_{X_iY_r}(t_1, t_2) - r_{X_rY_i}(t_1, t_2) \tag{3.102}$$

$$r_{X_rY_r}(t_1, t_2) = \int_{-\inf}^{+\inf} \int_{-\inf}^{+\inf} x_r y_r f_{Z_X Y(t_1, t_1')}(x_r, y_r, t_1, t_2) dx_r dy_r \tag{3.103}$$

$$r_{X_rY_r}(t_1, t_2) = \int_{-\inf}^{+\inf} \int_{-\inf}^{+\inf} x_i y_i f_{Z_X Y(t_1, t_1')}(x_i, y_i, t_1, t_2) dx_i dy_i \tag{3.104}$$

$$r_{X_rY_r}(t_1, t_2) = \int_{-\inf}^{+\inf} \int_{-\inf}^{+\inf} x_i y_r f_{Z_X Y(t_1, t_1')}(x_i, y_r, t_1, t_2) dx_i dy_r \tag{3.105}$$

$$r_{X_rY_r}(t_1, t_2) = \int_{-\inf}^{+\inf} \int_{-\inf}^{+\inf} x_r y_i f_{Z_X Y(t_1, t_1')}(x_r, y_i, t_1, t_2) dx_r dy_i \tag{3.106}$$

For real $X(t)$ and $Y(t)$, the cross-correlation is obtained by setting:

$$\begin{aligned} r_{X_iY_i}(t_1, t_2) =& \\ =& r_{X_iY_r}(t_1, t_2) \\ =& r_{X_rY_i}(t_1, t_2) \\ =& 0 \end{aligned} \tag{3.107}$$

in (1.100) as follows:

$$r_{XY}(t_1, t_2) = r_{X_rY_r}(t_1, t_2) \tag{3.108}$$

which is determined by (1.103).

3.4 Conclusions

In this chapter we have seen an overview of the mathematical foundations necessary in order to understand our method of studying the shape and classification of fibers. In particular we have described the properties to be verified by our metric and how similarity metrics (or dissimilarity) are applied to clustering algorithms.

We give a brief introduction explaining the metric spaces and the proprieties needed to talk about distance. Indeed, for talk about new distance is necessary to have the respect of to the distance properties. If we considering the fibers of the brain as a set of points in 3D space, that provide generic curves, we can study the behavior of fibers by means of differential geometry of curves and analyze the local and global aspect of them. The curves in the plane are described by intrinsic parameters that describe the geometry, for this reason the frame of Frénet has been presented that will be used as the first analysis of the behavior of the fibers. We have seen the limits of analysis if we consider the fibers in the plane and for this reason we are oriented towards the study of fibers in the three-dimensional case. To understand how much our new distance is a good distance, we have introduced the concept of correlation. The results of correlation is the similarity index that represents how many pairs of fibers are similar to each other.

In the next chapter we will see in detail how these mathematical aspects have been used to define the novel similarity metric introduced in this thesis.

4

The NewSimilarity metric

4.1 Introduction

Diffusion Imaging (DI) is a structural Magnetic Resonance Imaging (MRI) technique, which provides a non-invasive way to explore organization and integrity of the white matter structures in the human brain. DI information can be used in surgical planning and in the study of anatomical connectivity, brain changes and mental disorders.

From DI data, the white matter fiber tract can be reconstructed using a class of techniques called tractography. The reduced fiber set can be utilized:

- As a pre-processing stage for sophisticated algorithms that cannot deal with huge number of fibers, for example in atlas construction.
- To ease the computational burden, since the reduced set may be sufficient for detection of various diseases as long as the deduction is performed efficiently.
- As a tool for comparison since there might be many DI datasets of low resolution taken over the years, and a reduced set may help comparing new, high resolution acquisitions with the old ones as, for example, when an intra-subject follow-up is needed.

The notion of simplifying the dataset has been proposed in recent years and different frameworks have been suggested to address this issue. These frameworks are usually based on combination of clustering techniques and a distance measure.

In order to perform clustering, first a mathematical definition of fiber similarity (or, more commonly, fiber distance) must be specified. Then, pairwise fiber distance is to be calculated and used as input to a clustering algorithm.

The most common measures used for distance only capture the local relationship between streamlines but not the global structure of the fiber. Global structure, refer to the fiber variability shape. Together, local and global information, may define a good measure of similarity.

In order to provide such information, we developed a novel framework Figure(4.1) where both distance and shape information are considered during the clustering. In the following we summarize the mathematical model of our approach, we define its main features and we show its effectiveness by analyzing a test data set

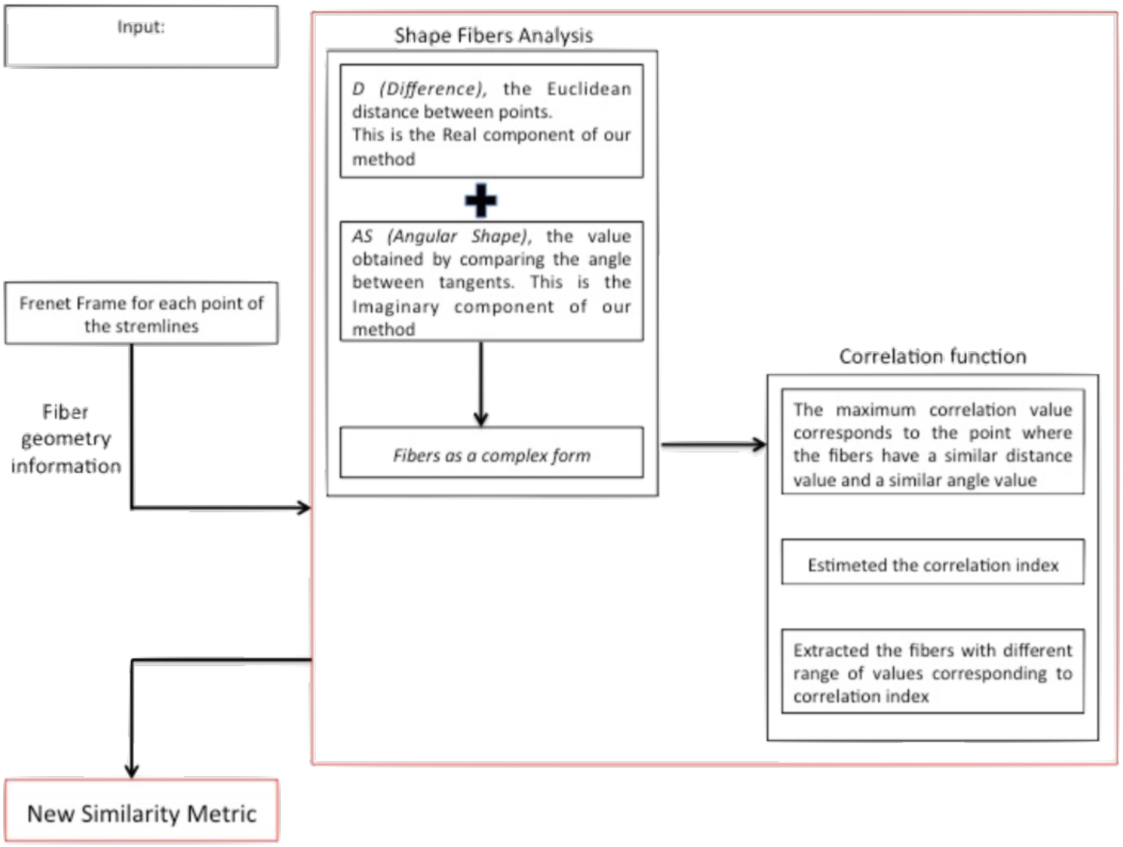

Input:
Frenet Frame for each point of the stremlines
Fiber geometry information
Shape Fibers Analysis
D (Difference), the Euclidean distance between points. This is the Real component of our method
AS (Angular Shape), the value obtained by comparing the angle between tangents. This is the Imaginary component of our method
Fibers as a complex form
Correlation function
The maximum correlation value corresponds to the point where the fibers have a similar distance value and a similar angle value
Estimeted the correlation index
Extracted the fibers with different range of values corresponding to correlation index
New Similarity Metric

Fig. 4.1: Framework of our method
Input consists of anatomical bundels. After the fiber geometry informations, with Frenet Frame, we computed a Shape Fibers Analysis and we obtained fibers as a complex form with introduction of two new parameters, Difference and Angular Shape. As the last step in exstracting our new similarity metric, we calculated the correlation index as an estimation of similarity between the fibers.

4.2 Mathematical Model

After tractography, fibers are extracted from DI images and represented as a set of streamlines in 3D space. Fiber tracking process starts from a set F of 3D polylines, each represented by a set of 3D points. To study the geometric organization of brain fibers, we introduce the following definitions.

Definition 4.1. *A fiber is a polyline defined on a sequence of p points in a* 3*-dimensional space,*
$f^i = \{x_1^i, ..., x_p^i\}$ *with* $x_m \in \mathbb{R}^3$, $m = 1, ..., p \in \mathbb{N}$, $i = 1, ..., n \in \mathbb{N}$.
We refer to a fiber also using the term streamline.

Definition 4.2. *Let a curve* $\gamma = (a, b) \to \mathbb{R}^3$ *such as* $\gamma(t_m) = x_m$ *with* $m = 1, ..., p \in \mathbb{N}$ *and* $a < t_1 <, ..., < t_p < b$.

We consider polylines as curves by parameterizing each polyline with respect to the arc-length and we use a parametric spline approximation to interpolate the curves.

Definition 4.3. *The set of n different 3D streamlines that belong to the same anatomical structure is called a track set or bundle*

$$F = \{\tilde{f}^1, ..., \tilde{f}^n\}.$$

Definition 4.4. *A metrics on a set X is a function (called the distance function or simply distance),*
$d : X \times X \mapsto \mathbb{R}$, where for all $x, y, z \in X$, the following conditions are satisfied:

- $d(x, y) \geq 0$ *non-negativity or separation axiom*
- $d(x, y) = 0 \Leftrightarrow x = y$ *identity of indiscernibles*
- $d(x, y) = d(y, x)$ *symmetry*
- $d(x, z) \leq d(x, y) + d(y, z)$ *triangle inequality*

In our case we define a measure of similarity between pairs of fibers, $\tilde{f}^i(t_m)$ and $\tilde{f}^j(t_m)$:

$$\begin{aligned} & \{\tilde{f}^i, \tilde{f}^j\} \in F, \\ \text{with} \quad & i, j = 1, ..., n \\ & i \neq j \\ & m = 1, ..., p \end{aligned} \tag{4.1}$$

4.2.1 Frenet-Frame

We analyze the fiber shape by means of the Frenet Frame representation of the curves $\tilde{f}$, which provides a flexible coordinate system determined by the geometric features of the curve itself.

A curve in space can be defined by a sequence of points, each representing a vector from the origin of a global X, Y, Z coordinate system. The first geometrical information of a curve $\tilde{f}$ is its length defined as:

$$L_{\tilde{f}} = \int_0^M ||\dot{\tilde{f}}||dt, \quad \text{where}|| \cdot || \quad \text{is the Euclidean norm in} \quad \mathbb{R}^3. \tag{4.2}$$

A fundamental result of the differential geometry of curves states that the arc-length:

$$s : t \mapsto s(t) = \int_0^t ||\dot{u}||du \quad \text{is an admissible parametrization of curves in} \quad F. \tag{4.3}$$

A direct consequence of the arc-length parametrization is that the derivative of the curve is the normalized tangent to the curve at point s, $T(s)$:

$$\dot{\tilde{f}}(t) = \dot{s}(t)T(s(t)) \tag{4.4}$$

The normalized curve can be completely characterized by introducing the Normal vector $N(s) = T^{'}(s)/||T'(s)||$, and the Bi-Normal vector $B(s) = T(s) \times N(s)$. The triplet $(T(s), N(s), B(s))$ forms the orthonormal Frenet-Serret frame and represents the local change in geometry of the curve. The shape of the curves can be retrieved from the evolution of the frame. Using Frenet frame it is possible to define two parameters intrinsic to the curve, Curvature and Torsion; in case of brain streamlines the torsion is not relevant since we do not consider the case of a fiber that go around itself.

The curvature could be a candidate parameter but 3D curvature is always non-negative whereas the 2D case allows for mixed-sign curvatures. Thus for 2D curves, the notion of "inflection point" (sign change) makes sense, but it does not for 3D curves.

Therefore, in the case of brain fibers that are curves in 3D space, analyzing the curvature allows us to distinguish how many fibers have similar shape. Only this information does not locate their positions in the anatomical bundle, as previously stated, we do not have 3D information on the sign and therefore on the bending direction. After having explained the parameterization that we have used and how we have created the pairs of points between the fibers, we will show the new parameters that have allowed us to overcome this problem.

Arc-length Parametrization

For to study the fibers bundles behavior composed by multiple 3d curves, we utilize a cosine series representation. The coordinates of curves are parameterized as coefficients of cosine series expansion.

Fourier descriptors have been used, in the past, to model planar curves and to classify of the shape of fiber tracts. These coefficients are computed by the Fourier transform that involves the both sine and cosine series expansion. In statistical and computational methods in brain image analysis, instead of using both the Fourier coefficients, sine and cosine, to obtain local informations on the shape of the fibers, it has been used only cosine representation. Unlike traditional splines, the proposed method does not have internal knots and explicitly represents curves as a linear combination of cosine basis.

We parameterize the arc-length as a unit interval. The $i-th$ control point in the fibers tract is mapped between 0 and 1. If we have the following streamline in 3D space, Figure(4.2), the parameterization of the tract is show in Figure(4.3). We can see the representation of cosine series at various degrees, Figure(4.4).

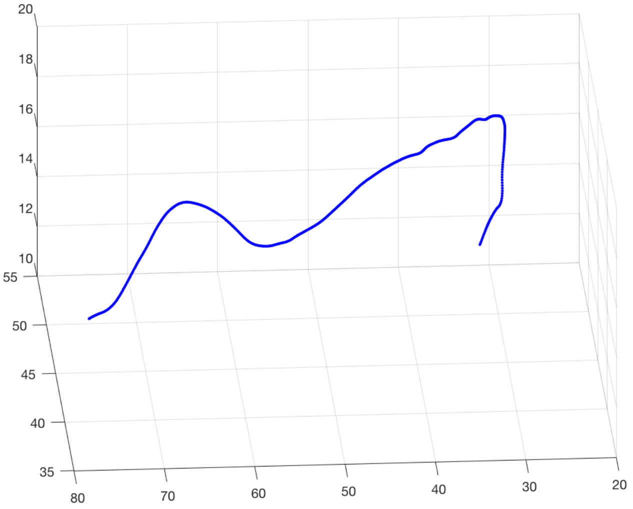


Fig. 4.2: Streamline in 3D space
Generic curve representing a possible brain fiber. Data obtained through the Matlab software.

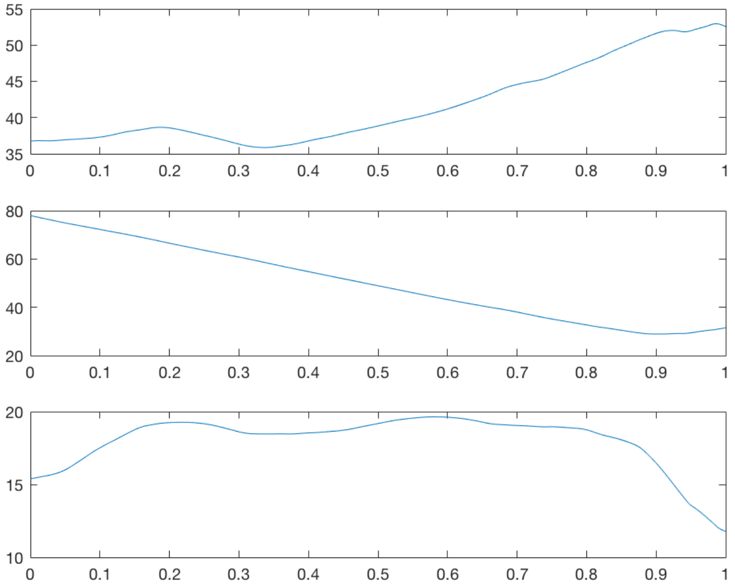


Fig. 4.3: Parameterization of the tract
Plots of x,y,z coordinates (axis of the ordinates) of original data and values of parameterization (axis of the abscissa), that is a number between 0 and 1.

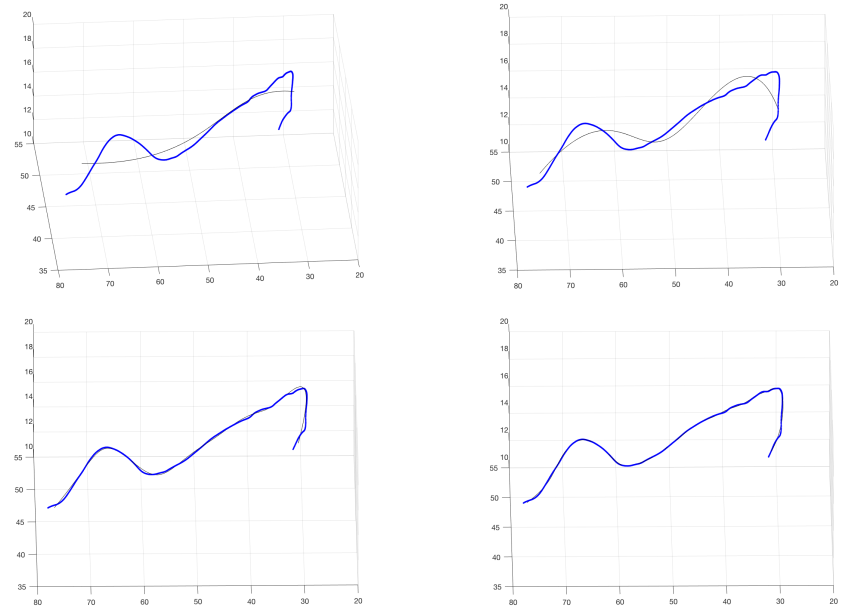


Fig. 4.4: The cosine series representation at various degrees
Blue lines are the original tract, while the black lines are the coordinates of cosine series representation. Starting from the left, we have the cosine series representation with one degree, five degree, ten degree and nineteen degree. Increasing the degree also increases the accuracy of the tract reconstruction.

Pairs of Points: Dynamic Time Warping

Two fibers may have different lengths and consequently different numbers of points in space; to quantify similarity between two fibers we use Dynamic Time Warping (DTW), that is a technique that looks for the optimal alignment of two sequence.

Consider two sequence of points,

$$\begin{aligned} &A = a_{1\times m}, ..., a_{i\times m}, ..., a_{p\times m} \\ &B = b_{1\times m}, ..., b_{i\times m}, ..., b_{p\times m} \\ &\text{with} \quad p \times m \quad \text{and} \quad q \times m \quad \text{dimension} \end{aligned} \tag{4.5}$$

where p and q refer to length of sequences and m represent the number of features. Two sequences can be arranged as a $p \times q$ matrix of the sides of a grid in which the distance between every possible combination of time instances a_i and b_j is stored. To find the best match between two sequences a path through the grid that minimizes the over all distance is needed. This path can be efficiently found using dynamic programming as follows:

$$d(a_i, b_j) = d_{EUC}(a_i, b_j) + min \begin{cases} d(a_{i-1}, b_j) \\ d(a_i, b_{j-1}) \\ d(a_{i-1}, b_{j-1}) \end{cases} \tag{4.6}$$

where $d_{EUC}(a_i, b_j)$ is the Euclidean distance between the $i - th$ point of sequence A and $j - th$ point of sequence B which can be calculated as:

$$d_{EUC}(a_i, b_j) = \sqrt{\sum_{l=1}^{m} (a_{i,l} - b_{j,l})^2} \tag{4.7}$$

Therefore overall Dynamic Time Warping distance between two sequences is:

$$DTW(A, B) = d(a_p, b_q) \tag{4.8}$$

In general, DTW is a method that calculates an optimal match between two given sequences with certain restrictions.

In this way, given a set of streamlines formed by a number of ordered points, we can find the best link between the points of fiber pairs. Ordered points mans that, with respect to anatomical bundles, we decide that fibers within the same bundles have the same initial and final point. Subsequently, between each pair of points, we calculated the parameters which compose our new similarity metric.

In order to also have information on the position and more extensive shape analysis of the fibers, we introduce the following new parameters for the analysis.

4.3 Method

We introduce a new distance measure between streamlines:

Definition 4.5. *Let $s : F \times F \mapsto \mathbb{C}$ be a function of distance between streamlines:*

$$s : (\tilde{f}^i(t_m), \tilde{f}^j(t_m)) \mapsto (D^{i,j}_m, \theta^{i,j}_m) \tag{4.9}$$

for each $\tilde{f}^i$ and $\tilde{f}^j$, with $\{\tilde{f}^i, \tilde{f}^j\} \in F$

$i, j = 1, ..., n$ $i \neq j$ and $m = 1, ..., p$

D represents the linear distance between points of two streamlines and θ represents the angular shape between tangents at corresponding points of two streamlines,where:

Definition 4.6. *Let $\{\tilde{f}^i, \tilde{f}^j\} \in F$ the first streamline with a sequence of p points $x_m \in \mathbb{R}^3$, and any of the other streamlines with the same number of points. We compute the Distance (D):*

$D^{i,j}_m : (\tilde{f}^i(t_m), \tilde{f}^j(t_m)) \mapsto \mathbb{R}$

$$D^{i,j}_m(\tilde{f}^i(t_m), \tilde{f}^j(t_m)) = ||x^i_m - x^j_m|| \tag{4.10}$$

for $m = 1, ..., p$

that contains the point-to-point distance between the first and the j-th fiber.

Definition 4.7. *Let $\{\tilde{f}^i, \tilde{f}^j\} \in F$ the first streamline with a sequence of p points in 3D space , and any of the remaining streamlines with the same number of points. We compute the Angular Shape (θ):*

$\theta^{i,j}_m : (\tilde{f}^i(t_m), \tilde{f}^j(t_m)) \mapsto \mathbb{I}$

$$\theta^{i,j}_m(\tilde{f}^i(t_m), \tilde{f}^j(t_m)) = cos^{-1}(\tilde{f}^{T^i}_{x_m} \cdot \tilde{f}^{T^j}_{x_m}) \tag{4.11}$$

where $\tilde{f}^{T^i}_m$ is the value at x_m of the tangent vector.

To analyze the shape of fibers, we have to introduce the above parameters that describe the new geometric distance computed, according to *Definition 1.5*, for each pair of fibers.

The values of the tangents are obtained from the Frenet frame previously calculated. In order to obtain even spatial information on the fiber pattern, we have calculated the internal product of the tangents values. The distance between the fibers is calculated from the standard Euclidean definition. The point coordinates refer to the coordinates of the data acquisition system, in our case to the magnetic resonance system.

For each fiber in the bundle, $\tilde{f} \in F$, the similarity is calculated for each pairs of points in two different fibers.

4.3.1 Similarity of fibers

Two fibers belong to the same bundle, if they have little distance separating them and if they have similar shapes. The criterion used to classify fibers is based on the correlation concept. In signal processing, correlation is a measure of similarity of two series as a function of the displacement of one relative to the other.

In our case, we have brain fibers with discrete sequences of points in 3D space. We apply discrete correlation, $(f \star g)[n] = \sum_{m=-\infty}^{+\infty} f^*[m]g[m+n]$, to each pair of fibers and obtain an index of maximum correlation, utilized for clustering the fibers. We introduce a new concept of similarity between fibers as a complex number, where the Real component composed by Euclidean distance and Imaginary component, composed by angular difference.
For combining this information, we use the polar coordinate reference system that defined a new similarity metric in the space of complex numbers.
These descriptions allow us to have a measure of both distance between the fibers and orientation of the tangent vectors.
We have grouped into a single equation, the distance *Definition 1.6*, and shape information of the fibers *Definition 1.7*:

Definition 4.8. *Given* $s : F \times F \mapsto \mathbb{C}$ *a distance between streamlines, then the* $NewSimilarity$ *is:*

$NewSimilarity = \sigma : \mathbb{C} \times \mathbb{C} \mapsto \mathbb{R}$

$$\sigma : ((D_m^{i,j}, \theta_m^{i,j}), (D_m^{i,j+1}, \theta_m^{i,j+1})) \mapsto \mathbb{R} \tag{4.12}$$

$((D_m^{i,j}, \theta_m^{i,j}), (D_m^{i,j+1}, \theta_m^{i,j+1})) = \textit{Correlation coefficient between pairs of fibers}$

Using this definition, we have information on how two fibers are oriented with respect to each other; indeed, calculating the angle for each pair of points of two different fibers, we can estimate how much the fibers are comparable in shape.

By means of this new distance it is possible to measure the correlation between pair of fibers represented as:

$\tilde{f}^{i,j} \mapsto (D_1^{i,j}, \theta_1^{i,j}), (D_2^{i,j}, \theta_2^{i,j}), ...(D_p^{i,j}, \theta_p^{i,j})$

$\tilde{f}^{i,j+1} \mapsto (D_1^{i,j}, \theta_1^{i,j}), (D_2^{i,j}, \theta_2^{i,j}), ...(D_p^{i,j}, \theta_p^{i,j})$

where $\tilde{f}^{i,j}$ represents the pair of fibers, in our case $\tilde{f}^{1,2}$, which will correlate with the rest of pairs in the dataset, $\tilde{f}^{i,j+1}$ $(\tilde{f}^{1,3})$. As mentioned above, the results of correlation is the similarity index that represents how many pairs of fibers are similar to each other. The index is expressed in complex form; we consider normalizing values to get a unitary index.

Similarity measures can be transformed into distance simply by taking the complement to 1. In this case, however, the "dissimilarity" term is preferred, respect to the distance term. The properties that must be met because a distance or dissimilarity coefficient is of metric type are those given in *Definition 1.4*.

In general, it is the fourth property the most important; the fact that it is or is not satisfied distinguishes metric measurements from those so-called semi-metric. Cross-Correlation has properties similar to an inner product, and can be used intuitively as a similarity function.

4.4 Method Performance

Below we report the behavior of synthetic data of the similarity metric.

In Figure 4.5, in red color, are highlighted the range where the fibers have a high correlation value. The set of values plotted (red and blu) are the information about the Real and Imaginary part like from *Definition 1.8*.

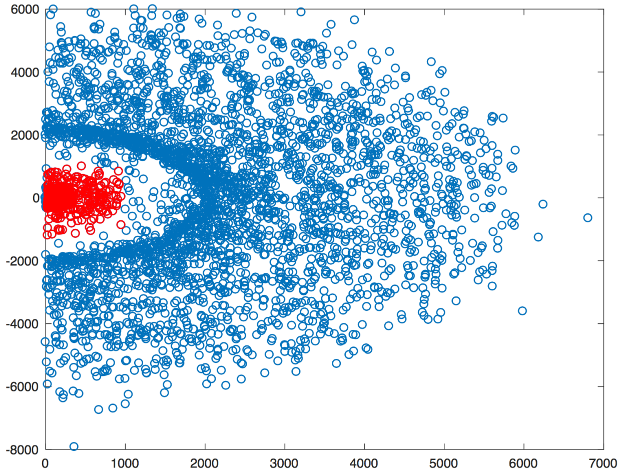


Fig. 4.5: Range of Fibers with high correlation value (red area)

Figure 4.6 shows the selected fibers respect to high correlation index (values in red color Figure 4.5). In black color, we have the reference fiber used for comparison with other fibers.

Otherwise, if we consider another range of values Figure 4.7, using the method, fibers in Figure 4.8 are selected, respect to high correlation index (values in red color Figure 4.7).

Compared to the reference fiber, in black color, the selected fibers are concentrated in the left part. These fibers have a low correlation value.

The Figure 4.9 and in Figure 4.10 shows the ranges, relating to high and low correlations between the fibers, respectively.

Fibers selected by these two ranges are shows in Figure 4.11 and in Figure 4.12. In black color the reference fiber.

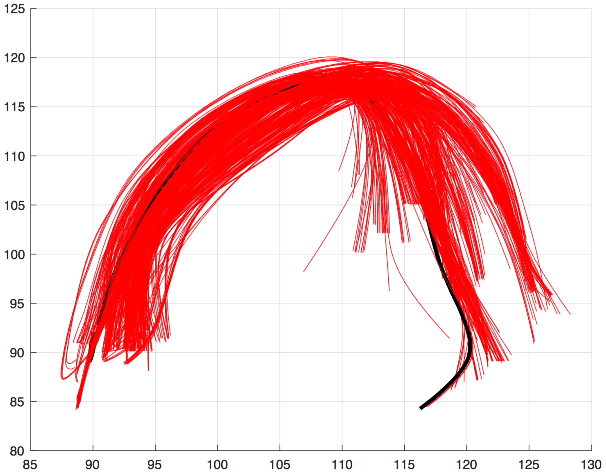


Fig. 4.6: Results of Fibers selected by the red area

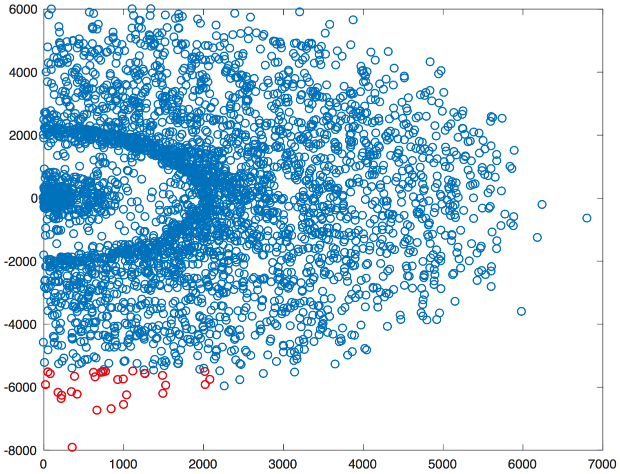


Fig. 4.7: Range of Fibers with low correlation value (red area)

In Figure 4.11 we have the fibers that are similar to each other, as far as concern distances and angles; instead in Figure 4.12 the fibers are less similar to each other.

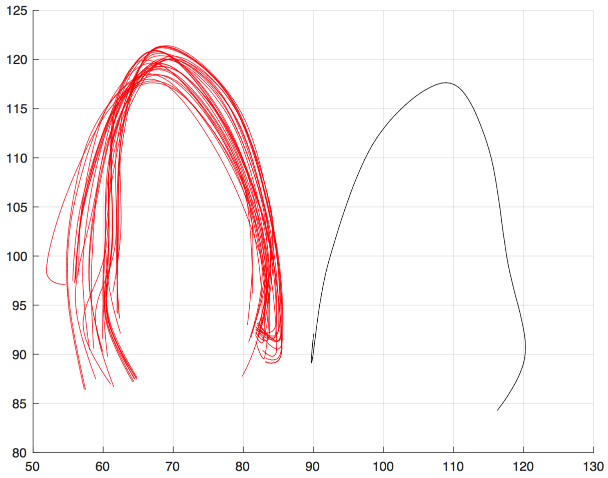


Fig. 4.8: Results of Fibers selected by the red area

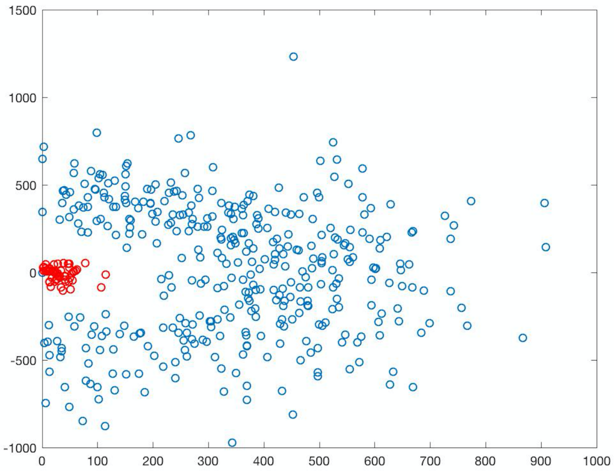


Fig. 4.9: Range of Fibers with high correlation value (red area)

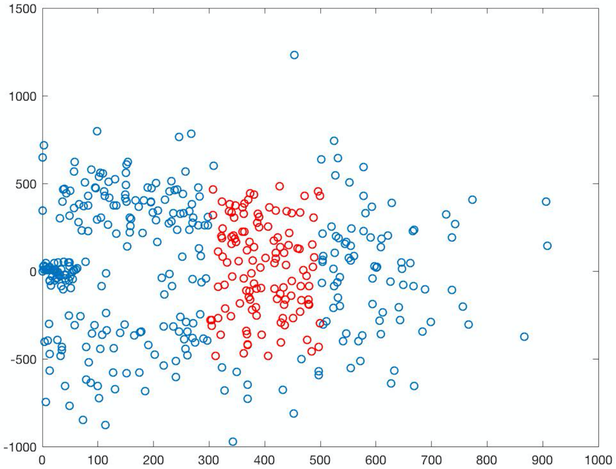


Fig. 4.10: Range of Fibers with low correlation value (red area)

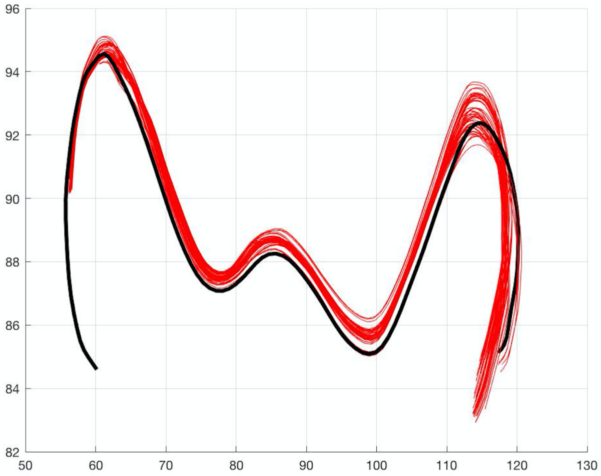


Fig. 4.11: Results of Fibers selected by the red area

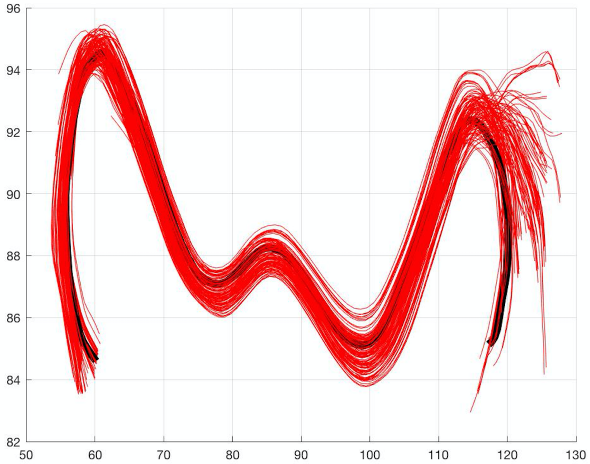


Fig. 4.12: Results of Fibers selected by the red area

4.5 Conclusions

The proposed method allows to classify the fibers on the basis of the correlation concept. In order to evaluate the goodness of the results we have selected two examples of values for each different set of fibers; a set where fibers appear more concentrated (Figure 2 and Figure 7), and a set where fibers appear more disparate (Figure 4 and Figure 8).

These figures provide a description of how, with respect to correlation values, fibers are placed in space. Each blue circle contains information about the proposal metric, *Definition 1.7* and *Definition 1.8*, of every pair of related fibers. In red circles, we have the informations extracted by correlation index. We considered these examples to see if the metric can distinguish the fibers that make up the beam with a distance and a similar angle between them and those with the values of these two parameters are very different from the fiber chosen as a reference.

As is shown in Figure 4.6, 4.8, 4.11 and 4.12 respect to the black fiber, that is the fiber of reference, the fibers are subdivided with respect to the similarity criterion. The results confirm a good fiber classification. In the next capther we want to consider synthetic anatomical dataset and we want to compare our metric with others and validate the clustering results by means of a ground-truth.

5

Experimental

5.1 Introduction

In this section we present the results of using the new similarity metrics, on datasets available online.

As the first step we analyze the characteristics of anatomical datasets used for verification, there areavailable online by ISMRM 2015 Tractography challenge - Data [48] .

The ISMRM 2015 Tractography challenge was based on an artificial phantom generated using the Fiberfox software [28], based on bundles segmented from a Human Connectome Project (HCP) subject. The aim was to create a realistic, clinical-style dataset that provided challenging bundles configurations. The challenge datasets are the ground-truth bundles for the generation of the dataset. The white matter bundles were manually segmentated based on definitions found in Diffusion Tensor Imaging, Introduction and Atla [59], which was written by challenge coorganizers Bram Stieltjes and Klaus Maier-Hein, as well as R.M. Brunner and F.B. Laun. Dataset are available with extension *.trk* and *.vtk*; this gives us the ability to use different software packages to view the data and gives us the ability to use the data both in the Matlab and in the Python environments. To use the union of different ground-truths we first loaded the file with the TrackVis software and then saved the dataset in the same file extension. As we will see in the next sections we have only used some of the ground-thruth. We chose the best known anatomical bundles that have geometric characteristics suitable for our analysis.

Afterwards we will show in a table the results obtained by applying the selected metrics to the datasets and how clusters occur in the different cases. We will present in detail the most significant cases. We use the Python library *scikit-learn* [57] for clustering classification.

5.2 Datasets

Figure(5.1), Figure(5.2) and Figure(5.3) show the three main datasets used in our processing.

The cingulum (CA) Figure(5.1) that is a collection of white matter fibers projecting from the cingulate gyrus to the entorhinal cortex in the brain, allowing for communication between components of the limbic system.

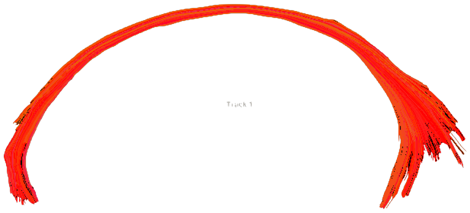


Fig. 5.1: CA bundle

The corticospinal tract (CST) Figure(5.2) is a white matter motor pathway starting at the cortex that terminates on motor neurons and interneurons in the spinal cord, controlling movements of the limbs and trunk.

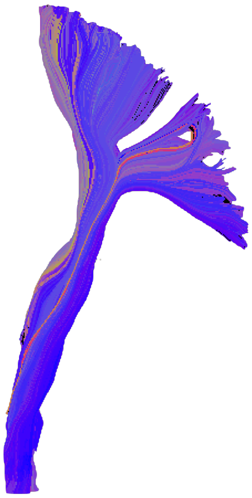


Fig. 5.2: CST bundle

The Fornix Figure(5.3) is a C-shaped bundle of nerve fibers in the brain that acts as the major output tract of the hippocampus. The fibers begin in the hippocampus on each side of the brain as fimbriae; the separate left and right sides are each called the crus of the Fornix. The bundles of fibers come together in the midline of the brain, forming the body of the fornix.

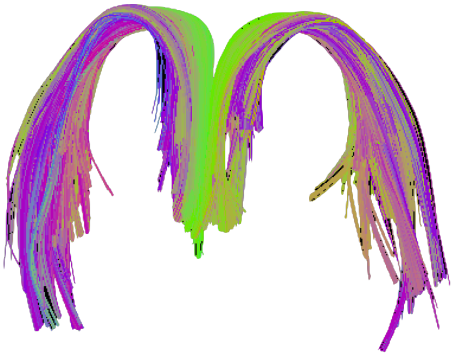


Fig. 5.3: Fornix bundle

We consider the case where two anatomical bundles are present, and for a more extensive application of the metrics we consider both to have two close anatomical bundles, are to be spaced apart from each other as shown in Figure(5.4).

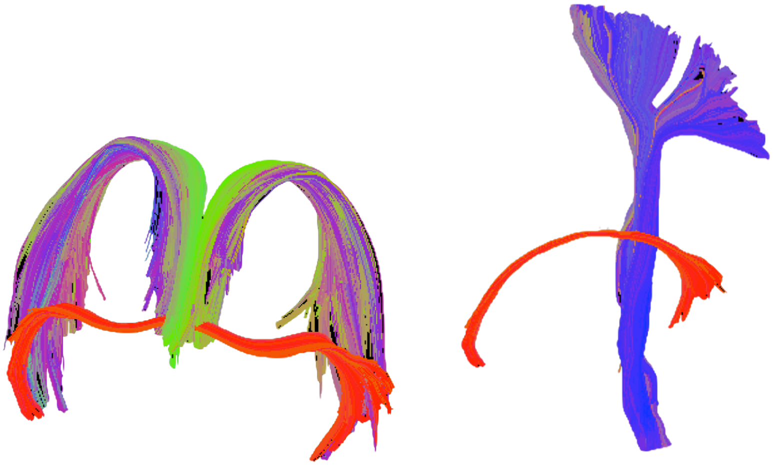


Fig. 5.4: Bundles of streamlines where fibers are spatial near (left) and far spatial (right) between them.

5.2.1 Analysis

In the analysis phase of the selected fiber bundles, we applied to the fibers our new metrics and other distance metrics most used in the literature; the results obtained through the clustering algorithms, have been qualitatively evaluated by means of typical indices, as we will see later in the summary tables.

The indices used need, as input of the algorithm, a ground truth of the fibers, i.e. the labels that identify the real position of the fibers inside the anatomical bundle. The ground truth we used was that of dividing the labels according to the bundles considered; specifically, considering the dataset composed of the fornix bundle and the CA bundle we obtained the number of fibers of the two bundles and assigned the value 0 to the 3831 fibers of the fornix bundle and value 1 to the 431 fibers of the CA bundle. In the case of the dataset composed by the bundles CST and CA we have given label value 0 for the fibers of the CST bundles and value 1 for the second bundles. We have analyzed the Fornix bundle alone to show how our metrics correctly identifies the orientation of the fibers relative to the other metrics. The orientation of the fibers indicates the fibers that are on the left side with respect to those on the right side. Since we do not have an anatomical ground truth of the fibers of this bundle we limit ourselves to show only visually the results obtained.

In our analysis we consider only some of the metric described in the previous chapters:

- Frechet metric
- Hausdorff metric
- MDF metric
- and the our metrics that we call "NewSimilarity metric"

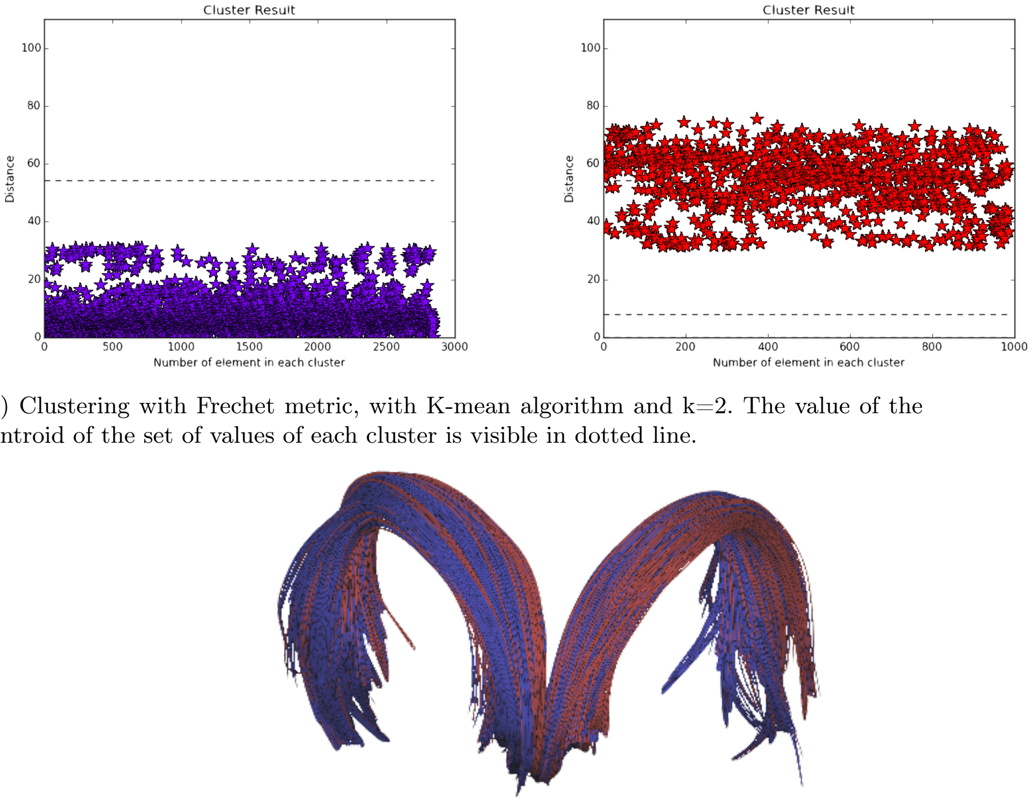
The steps used for in this analysis are summarized below.

- Choice of input dataset: the image of the initial dataset gives us information on the type of data we want to analyze; in this case the fibers are colorless because we have yet to start the analysis of the metrics.
- Quantitative analysis of the dataset. The similarity metrics are applied to the initial dataset and a clustering algorithm is chosen for the classification of the fibers. Consider two graphs to represent this analysis; the first one where the fibers of the initial dataset are colored after the clustering. In this way we have a first visual result of the classification of the fibers. The second graph, where in x-axis we represent the value of similarity fibers and in y-axis we represent the number of element(fibers) in each cluster. Compared to the similarity value, we see the corresponding number of fibers classified within the cluster.
- Qualitative analysis of the dataset. To facilitate the comparison between the different metrics, we report a summary table of the clustering performance with the index values.

Input Dataset: Fornix

The first case we consider is the single anatomical bundle. This bundle is composed of two sides, left and right. The fibers have different lengths and different shapes

Figure(5.5a) shows the clustering results and the fibers colored with respect to the clustering algorithm with Frechet metric and K-mean algorithm; in red color the result of the first cluster and in violet color the result of the second cluster. On the x-axis we have the number of elements (fibers) and on the y-axis we have the value of distance. The value of the centroid of the set of values of each cluster is visible in dotted line. Figure(5.5b) shows the output of Frechet metric using different colors for the fiber bundles; in this case, the distance metric does not clearly distinguish the different orientations of the fibers.

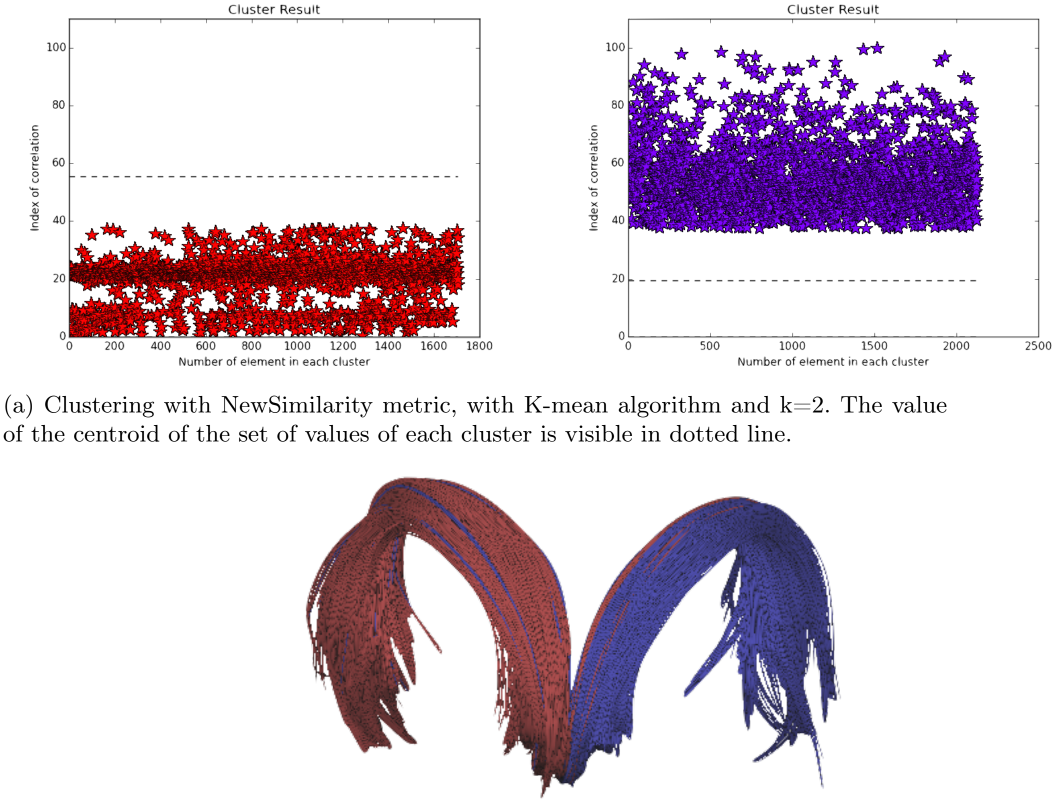


(a) Clustering with Frechet metric, with K-mean algorithm and k=2. The value of the centroid of the set of values of each cluster is visible in dotted line.

(b) Fornix Fibers Clustering with Frechet metric.

Fig. 5.5: Fornix dataset. Clustering with Frechet metric.

Figure(5.6a) shows the clustering result and the fibers colored as output of the clustering algorithm with NewSimilarity metric; in red color the result of the first cluster and in violet color the result of the second cluster. In the x-axis we have the number of elements (fibers) and in y-axis we have the index of correlation. The value of the centroid of the set of values of each cluster is visible in dotted line. Figure(5.6b) shows the output of the NewSimilarity metrics using different colors for the fiber bundles; in this case the division of the fibers with respect to the different orientations appears more marked. With our metrics we are able to give both a distance contribution and a form contribution to classify the fibers.



(a) Clustering with NewSimilarity metric, with K-mean algorithm and k=2. The value of the centroid of the set of values of each cluster is visible in dotted line.

(b) Fornix Fibers Clustering with NewSimilarity metric.

Fig. 5.6: Fornix dataset. Clustering with NewSimilarity metric.

The comparison between Figure(5.6b) and Figure(5.5b) shows that the fibers classified with the Frechet metric do not have a good distribution. The final classification of fibers appears disordered, with respect to our method. In this case, the only metric distance used in Frechet is not enough to classify the fibers. The form-based classification included in our method best suits this dataset.

Whith our metrics it is possible to have a separate fiber clustering. Only a restricted group of fibers are classified in the wrong part, as shown in Figure(5.6b). If we consider the left and right parts of Fornix bundle, we observe that at the left bundle, we have a few fibers of violet color and vice-versa if we look at the right bundle we have few red fibers. These fibers of different colors represent for us the fibers that are not classified correctly; in Figure(5.5b) the fibers classified with different colors are of a greater number

We considered the right or wrong part, without anatomical or surgical know-how. With this type of dataset we demonstrate that the metrics used is able to recognize and then to classify the fibers in two distinct sub groups. At a later stage, we will analyze the clustering of fibers by validating the result with respect to the brain anatomy.

Input Dataset: Fornix and CA

In this example Figure(5.7), we consider two anatomical bundles the Fornix bundle Figure(5.3) dataset is composed by fibers whit different length and orientation, and the CA bundle Figure(5.1) where fibers have the different length but similar direction and are spatially close to each other. Using two spatially close datasets, we want to see which parameter influences the most the classification of fibers; specifically, we are interested in understanding whether the metrics distinguishes the two different anatomic bundles: i.e. if it is sufficient to consider only the distance between the fibers or if it is necessary to introduce the shape parameter to improve the final result.

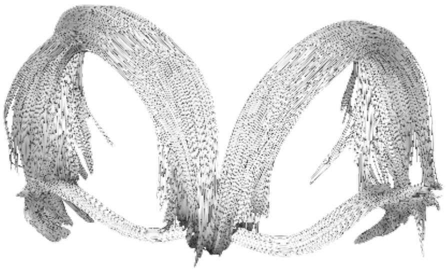
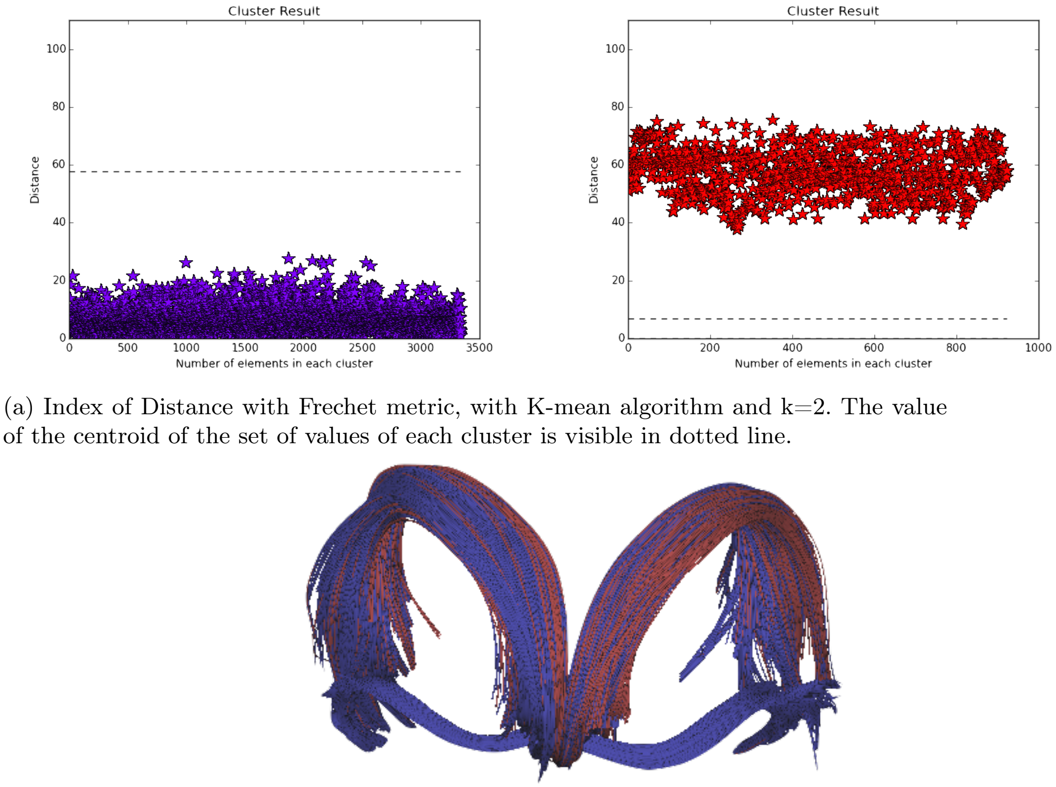


Fig. 5.7: Fornix and CA Dataset. The bundles are close to each other. Fornix dataset is composed of fibers whit different length and orientation. We considered it divided in two sides, right and left. The second dataset is CA, here the fibers have different lengths but similar direction.

Figure(5.8a) shows the distance index using the Frechet metric; in violet color the result of the first cluster and in red color the result of the second cluster. In x-axis we have the number of elements (fibers) and in y-axis we have the value of distance. The value of the centroid of the set of values of each cluster is visible in dotted line. Figure(5.8b) shows the classification of the fibers with K-mean algorithm and k=2; the final result shows a correct classification of the CA bundle, but we do not have a clear division of the Fornix bundle; only a few red fibers (in this case) are classified on the right side and many violet fibers are classified both on the right side and on the left side.

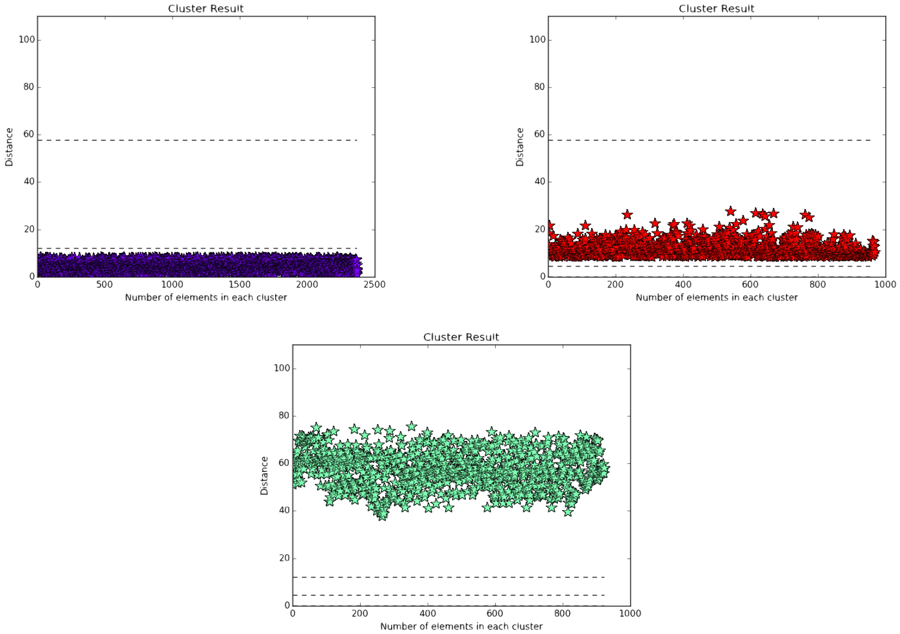


(a) Index of Distance with Frechet metric, with K-mean algorithm and k=2. The value of the centroid of the set of values of each cluster is visible in dotted line.

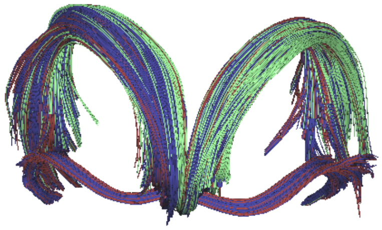
(b) Clustering result with Frechet metric.

Fig. 5.8: Fornix and CA dataset. Clustering with Frechet metric with K-mean algorithm and k=2.

Figure(5.9a) shows the distance index using the Frechet metric with K-mean algorithm and k=3; in violet color the result of the first cluster, in red color the result of the second cluster and in green color the result of the third cluster. In x-axis we have the number of elements (fibers) and in y-axis we have the value of distance. The value of the centroid of the set of values of each cluster is visible in dotted line. Figure(5.9b) shows that the result of clustering is not homogeneous; in this case, considering k = 3, we did not obtain a good classification of the CA bundle and also the classification of the fornix bundles seems more confused than the results with k = 2.



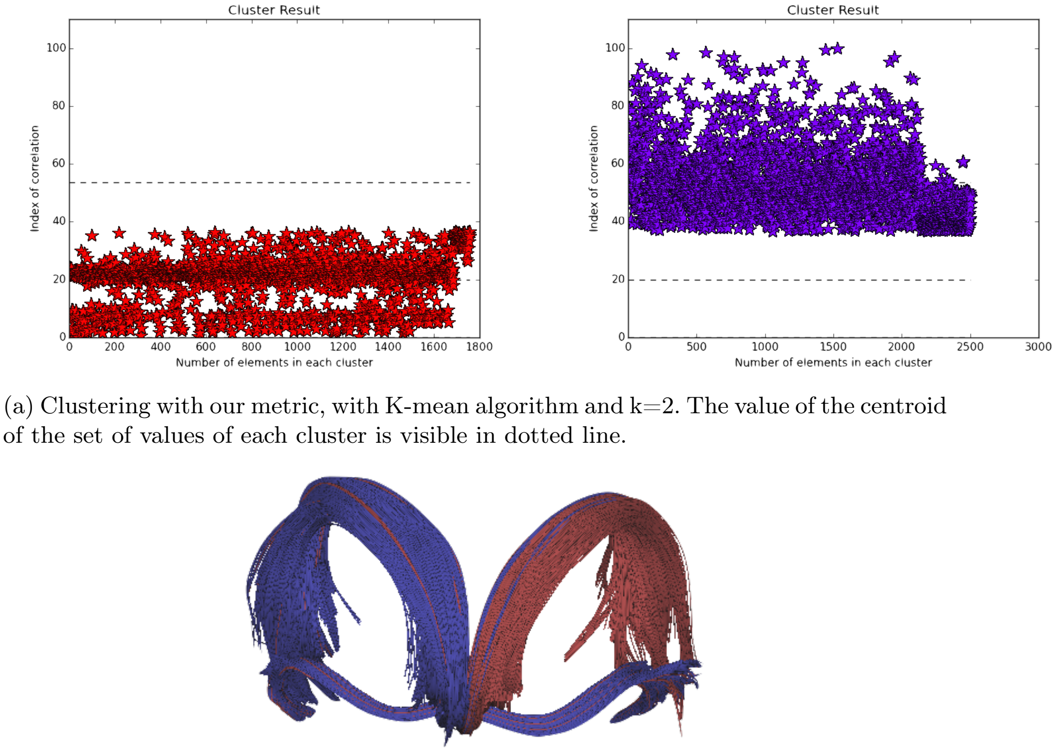
(a) Index of Distance with Frechet metric, with K-mean algorithm and k=3. The value of the centroid of the set of values of each cluster is visible in dotted line.



(b) Clustering result with Frechet metric, with K-mean algorithm and k=3.

Fig. 5.9: Fornix and CA dataset. Clustering with Frechet metric with K-mean algorithm and k=3.

Figure(5.10a) shows the correlation index using NewSimilarity metric; in red color the result of the first cluster and in violet color the result of the second cluster. In x-axis we have the index of correlation otherwise in y-axis we have the number of elements (fibers). Figure(5.10b) shows the classification of the fibers with K-mean algorithm and k=2; in this case we have a better division of the Fornix bundle using our metric, compared to previous results with different metrics. The CA bundle is classified together with the fibers of the fornix bundle of violet color.



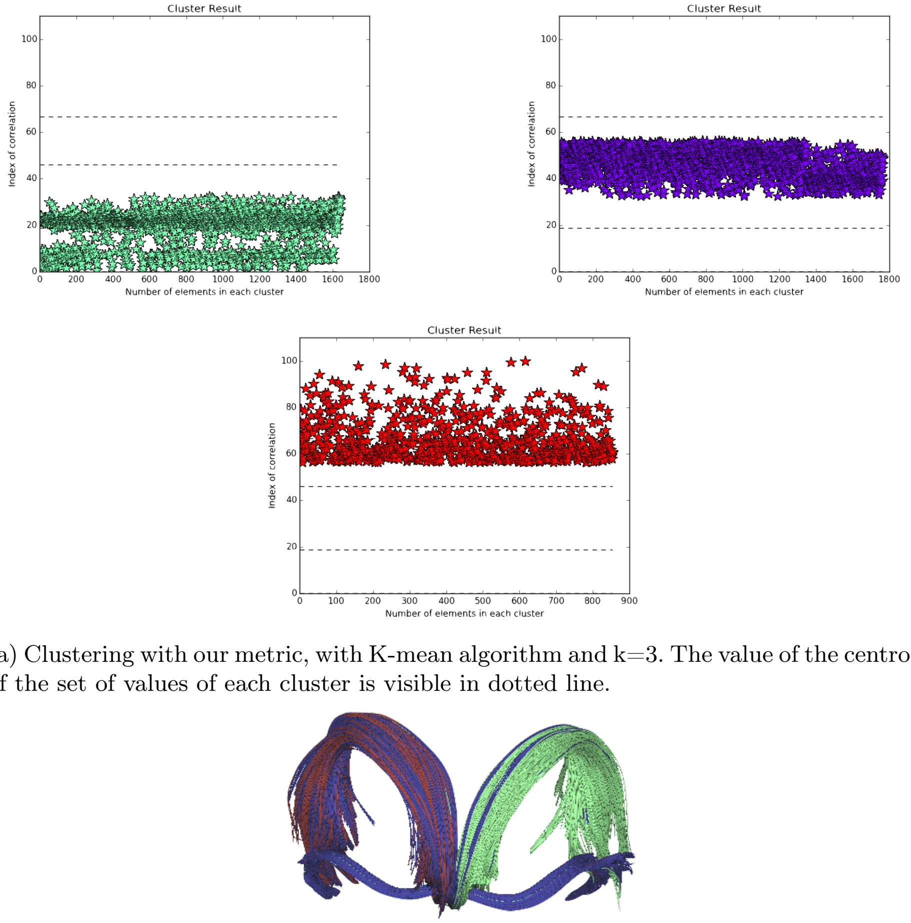
(a) Clustering with our metric, with K-mean algorithm and k=2. The value of the centroid of the set of values of each cluster is visible in dotted line.

(b) Fornix and CA Fibers Clustering with NewSimilarity metric, with K-mean algorithm and k=2.

Fig. 5.10: Fornix and CA dataset. Clustering with NewSimilarity metric with K-mean algorithm and k=2.

We get obtain the classification of Figure(5.10b) by considering two clusters input divides the left side and the right side of the Fornix bundle quite homogeneously, and the whole bundle of the same color (violet) is colored. The CA bundle is grouped together with part of the Fornix bundle, this result is due to the fact that considering in our method the pair of fibers the method finds a similarity in shape and proximity, violet color of the fibers; instead, we have red fibers at the time the fibers are distant from each other and the comparison of the pairs of fibers as a final result of the disomogeneities.

Figure(5.11a) shows the correlation index using our metric with K-mean algorithm and k=3; in green color the result of the first cluster, in violet color the result of the second cluster and in red color the third cluster. In x-axis we have the number of elements (fibers) and in y-axis we have the index of correlation. The value of the centroid of the set of values of each cluster is visible in dotted line. Figure(5.11b) shows the classification of the fibers with K-mean algorithm and k=3; in this case we have a better division of the Fornix bundle using our metric, compared to previous results with different metrics. With our metric, compared to the Frechet metric, the fibers are classified with better behavior. The metric identifies the 3 clusters in a clearer way. This result is obtained with respect to a value of positive ARI and low inertia.



(a) Clustering with our metric, with K-mean algorithm and k=3. The value of the centroid of the set of values of each cluster is visible in dotted line.

(b) Fornix and CA Fibers Clustering with NewSimilarity metric, with K-mean algorithm and k=3.

Fig. 5.11: Fornix and CA dataset. Clustering with NewSimilarity metric with K-mean algorithm and k=2.

Table(5.1), shows the tables that contain the indices to evaluate the clustering performance. The indices we have considered are:

- homogeneity score(homo): each cluster contains only members of a single class.
- completeness score(compl): all members of a given class are assigned to the same cluster
- V-masure(v-meas): entropy-based measure which explicitly measures how successfully the criteria of homogeneity and completeness have been satisfied.
- adjusted Rand index(ARI): computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings.
- adjusted mutual information(AMI): function that measures the agreement of the two assignments, ignoring permutations.
- silhouette coefficient(silhouette): measures how close a fiber is to its own cluster in comparison to the rest of the clusters, i.e. whether there is another cluster that might represent it better or as well. Silhouette analysis can be used to study the separation distance between the resulting clusters.

The clustering algorithm used is K-means and the analyzed metrics are Frechet metric, Hausdorff metric, MDF metric and NewSimilarity metric.

Figure(5.1a) shows clustering performance of Frechet metric. With this metric we obtain high values of inertia, values that refer to the measure of coherence of the clusters. We remember that low values are preferable for this index. Negative values of the Adjusted Rand index (ARI) indicate that we have independent labelings between them; indeed similar clusterings have a positive ARI. Both in the case of clusters=2 and clusters=3 the value of silhouette coefficient is good; range is between [-1, 1]. The rest of values are reasonable.

Figure(5.1b) shows clustering performance of Hausdorff metric. Compared to the previous metric we obtain lower values of inertia, values that refer to the measure of coherence of the clusters. We remember that low values are preferable for this index. Negative values of the Adjusted Rand index (ARI) indicate that we have independent labelings between them; indeed similar clusterings have a positive ARI. Both in the case of clusters=2 and clusters=3 the value of silhouette coefficient is good; range is between [-1, 1]. The rest of values are reasonable.

Figure(5.1c) shows clustering performance of MDF metric. As for the previous metric we obtain lower values of inertia, values that refer to the measure of coherence of the clusters. We remember that low values are preferable for this index. Negative values of the Adjusted Rand index (ARI) indicate that we have independent labelings between them; indeed similar clusterings have a positive ARI. Both in the case of clusters=2 and clusters=3 the value of silhouette coefficient is good; range is between [-1, 1]. The rest of values are reasonable.

Figure(5.1d) shows clustering performance of NewSimilarity metric. In this case we have low values of inertia. Both in the case of clusters=2 and clusters=3 the value of ARI index is positive, even if with a low value. We remember that low values are preferable for this index. Negative values of the Adjusted Rand index (ARI) indicate that we have independent labelings between them; indeed similar clusterings have a positive ARI. The values of silhouette coefficient is good, because

Metric: Frechet
Clustering Algorithm: K-means

Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	112098	0.076	0.048	0.058	-0.089	0.047	0.889
n_cluster 3	73843	0.077	0.025	0.038	-0.030	0.025	0.647

(a) Clustering Performance with Frechet metric.

Metric: Hausdorff
Clustering Algorithm: K-means

Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	70528	0.078	0.048	0.060	-0.088	0.048	0.831
n_cluster 3	34999	0.085	0.028	0.042	-0.037	0.027	0.643

(b) Clustering Performance with Hausdorff metric.

Metric: MDF
Clustering Algorithm: K-means

Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	93887	0.077	0.048	0.059	-0.089	0.048	0.779
n_cluster 3	50186	0.072	0.029	0.041	-0.083	0.029	0.670

(c) Clustering Performance with MDF metric.

Metric: NewSimilarity
Clustering Algorithm: K-means

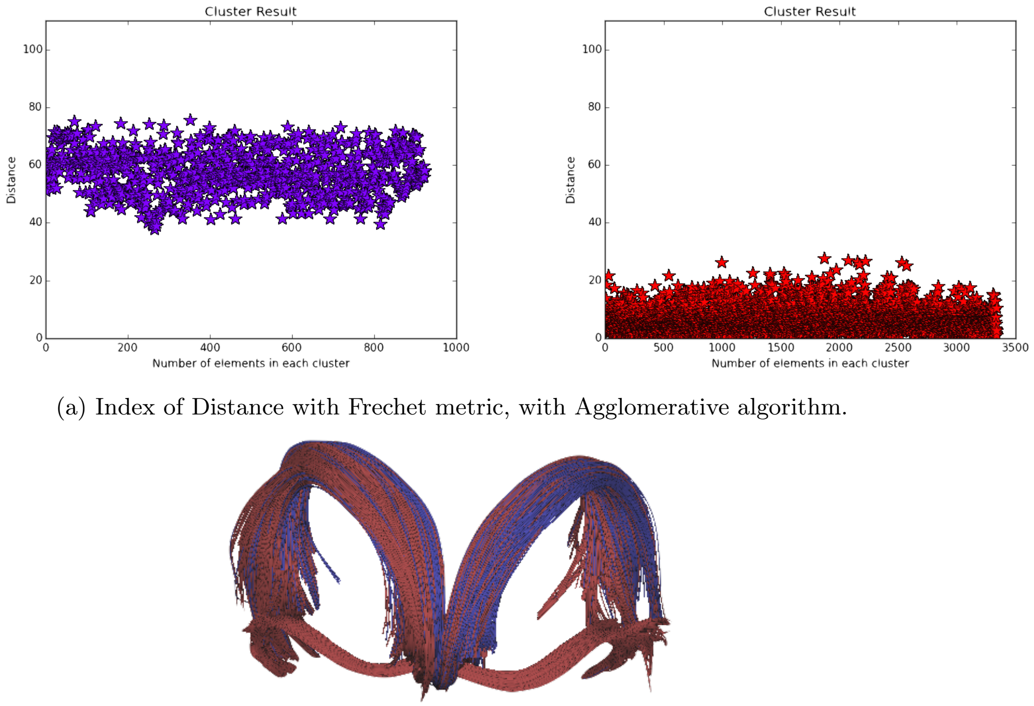
Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	46	0.048	0.023	0.032	0.018	0.023	0.646
n_cluster 3	22	0.239	0.074	0.113	0.004	0.074	0.610

(d) Clustering Performance with NewSimilarity metric.

Table 5.1: Clustering Performance using different metrics and K-means algorithm.

are equal to 0.6; these values are slightly lower than the other metrics. The rest of values are reasonable.

Figure(5.12a) shows the index of distance using Frechet metrics and the Agglomerative algorithm; in violet color the result of the first cluster and in red color the result of the second cluster. In x-axis we have the number of elements (fibers) and in y-axis we have the index of distance. Figure(5.12b) shows clustering result with Frechet metrics. The final result shows a correct classification of the CA bundle, but we do not have a clear division of the Fornix bundle; neither part (left or right) of the fornix bundle is classified together with the CA tract (in red color). This is the result we expected, in spite of considering k = 2 is an underestimation of clustering, forcing the result. This is the result we expected, despite considering k = 2 is an underestimation of clustering (we force the result).

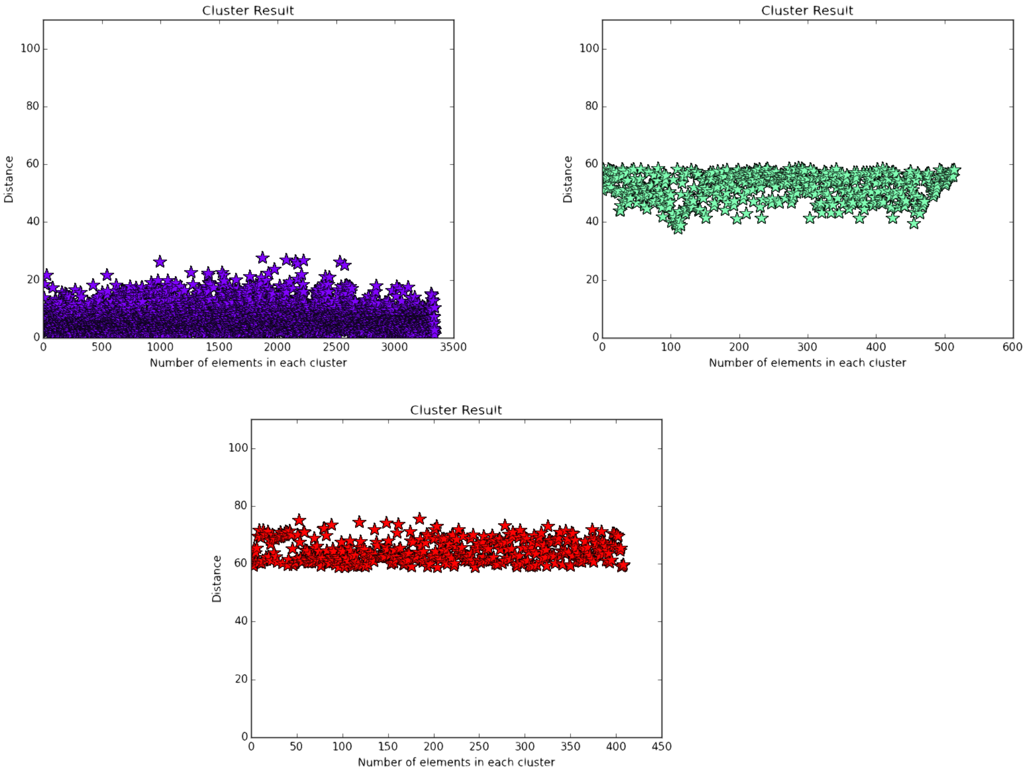


(a) Index of Distance with Frechet metric, with Agglomerative algorithm.

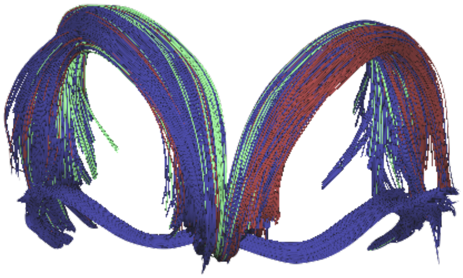
(b) Clustering result with Frechet metric.

Fig. 5.12: Fornix and CA dataset. Clustering with Frechet metric with Agglomerative algorithm (numbers of cluster are two).

Figure(5.13a) hows the index of distance using Frechet metrics and the Agglomerative algorithm; in violet color the result of the first cluster, in green color the result of the second cluster and in red color third cluster. In x-axis we have the number of elements (fibers) and in y-axis we have the index of distance. Figure(5.13b) shows clustering result with Frechet metric. The final result shows, with respect to the values obtained with K-mean algorithm, an bad fibers classification.



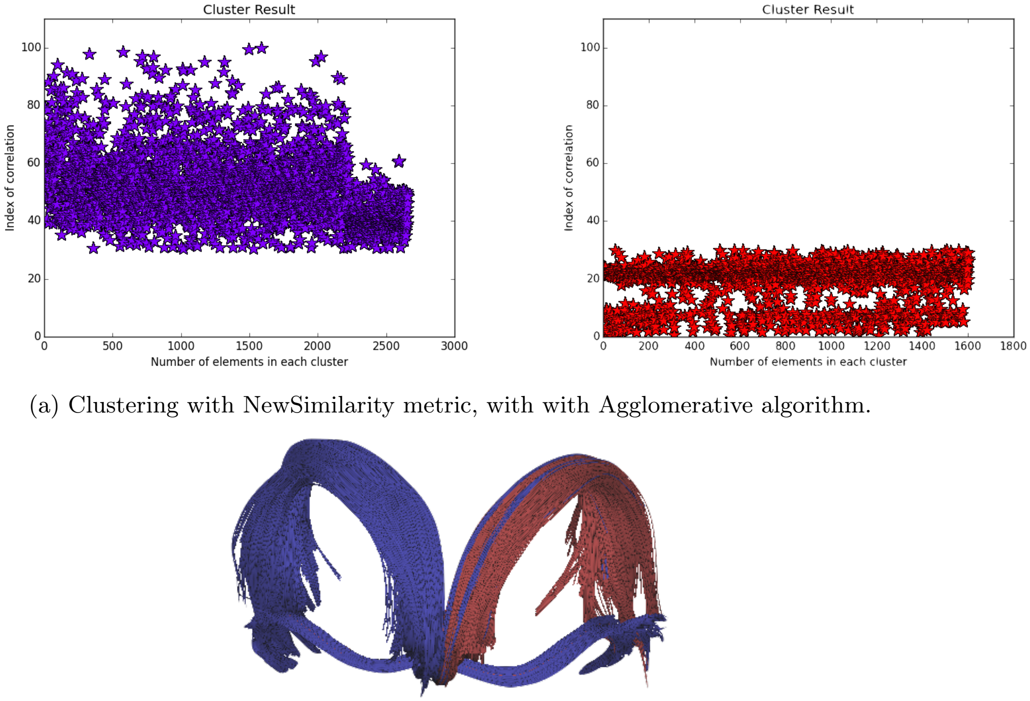
(a) Index of Distance with Frechet metric, with with Agglomerative algorithm.



(b) Clustering result with Frechet metric, with with Agglomerative algorithm.

Fig. 5.13: Fornix and CA dataset. Clustering with Frechet metric with Agglomerative algorithm (numbers of cluster are three).

Figure(5.14a) shows the index of distance using NewSimilarity metrics and the Agglomerative algorithm; in violet color the result of the first cluster and in red color the result of the second cluster. In x-axis we have the number of elements (fibers) and in y-axis we have the index of correlation. Figure(5.14b) shows clustering result with NewSimilarity metric. The final result shows a correct classification of the CA bundle, most of the fibers of the Fornix bundle are classified in a consistent manner with respect to the classification of the CA bundle. Compared to the Frechet metric using NewSimilarity metric we obtain, as shown in Figure(5.2), higher values of completeness and silhouette and we do not have negative values of the ARI index.



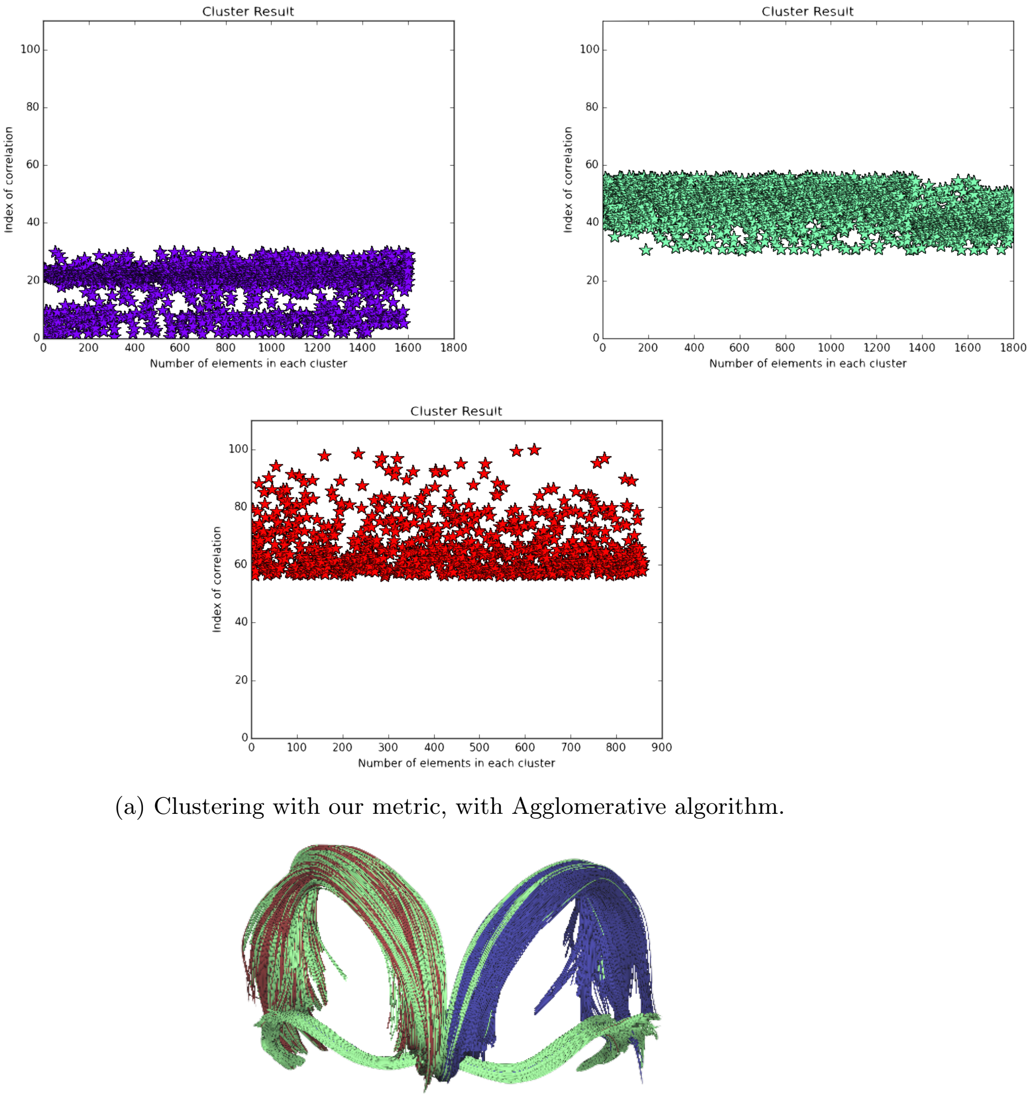
(a) Clustering with NewSimilarity metric, with with Agglomerative algorithm.

(b) Fornix and CA Fibers Clustering with NewSimilarity metric, with Agglomerative algorithm.

Fig. 5.14: Fornix and CA dataset. Clustering with NewSimilarity metric with Agglomerative algorithm (numbers of cluster are two).

Figure(5.15a) shows the index of distance using NewSimilarity metrics and the Agglomerative algorithm; in violet color the result of the first cluster, in green color the result of the second cluster and in red color the result of the third cluster. In x-axis we have the number of elements (fibers) and in y-axis we have the index of correlation. Figure(5.15b) shows clustering result with NewSimilarity metric.

The final result shows a better classification if we compare the results with those obtained using the Frechet metric.



(a) Clustering with our metric, with Agglomerative algorithm.

(b) Fornix and CA Fibers Clustering with our metric, with Agglomerative algorithm.

Fig. 5.15: Fornix and CA dataset. Clustering with NewSimilarity metric with Agglomerative algorithm (numbers of cluster are three).

Table(5.2), shows the tables that contain the index for evaluate the clustering performance. The algorithm used for fibers classification is Agglomerative clustering and the analyzed metrics are Frechet metric, Hausdorff metric, MDF metric and NewSimilarity metric.

Metric: Frechet
Clustering Algorithm: Agglomerative

Cluster	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	0	0.048	0.058	-0.089	0.047	0.634
n_cluster 3	0	0.037	0.050	-0.106	0.037	0.592

(a) Clustering Performance with Frechet metric.

Metric: Hausdorff
Clustering Algorithm: Agglomerative

Cluster	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	0	0.048	0.058	-0.089	0.047	0.829
n_cluster 3	0	0.029	0.044	-0.050	0.029	0.647

(b) Clustering Performance with Hausdorff metric.

Metric: MDF
Clustering Algorithm: Agglomerative

Cluster	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	0	0.048	0.059	-0.089	0.047	0.780
n_cluster 3	0	0.037	0.050	-0.106	0.037	0.722

(c) Clustering Performance with MDF metric.

Metric: NewSimilarity
Clustering Algorithm: Agglomerative

Cluster	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	0	0.060	0.009	0.008	0.006	0.889
n_cluster 3	0	0.090	0.028	0.030	0.019	0.825

(d) Clustering Performance with NewSimilarity metric.

Table 5.2: Clustering Performance using different metrics and Agglomerative algorithm.

Table(5.2a) shows clustering performance of Frechet metric. We use the agglomerative algorithm we do not have the index of inertia because this index is representative of the k-means algorithm. We have high values of the silhouette coefficient for both clusters. The values of the ARI index are negative as in the case of k-means algorithm.

Table(5.2b) shows clustering performance of Hausdorff metric. We use the agglomerative algorithm we do not have the index of inertia because this index is representative of the k-means algorithm. We have high values of the silhouette coefficient for both clusters. The values of the ARI index are negative as in the case of k-means algorithm.

Table(5.2c) shows clustering performance of MDF metric. We use the agglomerative algorithm we do not have the index of inertia because this index is representative of the k-means algorithm. We have values similar to the other metrics.

Table(5.2d) shows clustering performance of Similarity metric. We use the agglomerative algorithm we do not have the index of inertia because this index is representative of the k-means algorithm. With our metric we have, respect k-means algorithm, negative values of ARI index.

Input Dataset: CST and CA

In this example Figure(5.16), we consider two anatomical bundles, CST bundle Figure(5.2) dataset is composed by fibers whit different length and orientation, and CA bundle Figure(5.1) where fibers have the different length but similar direction. This anatomical bundles has fibers that are spatially distant from each other.

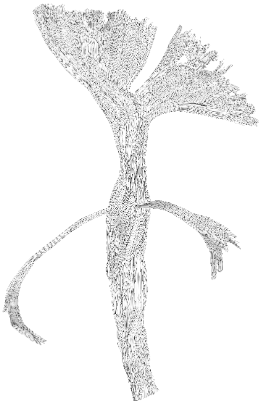
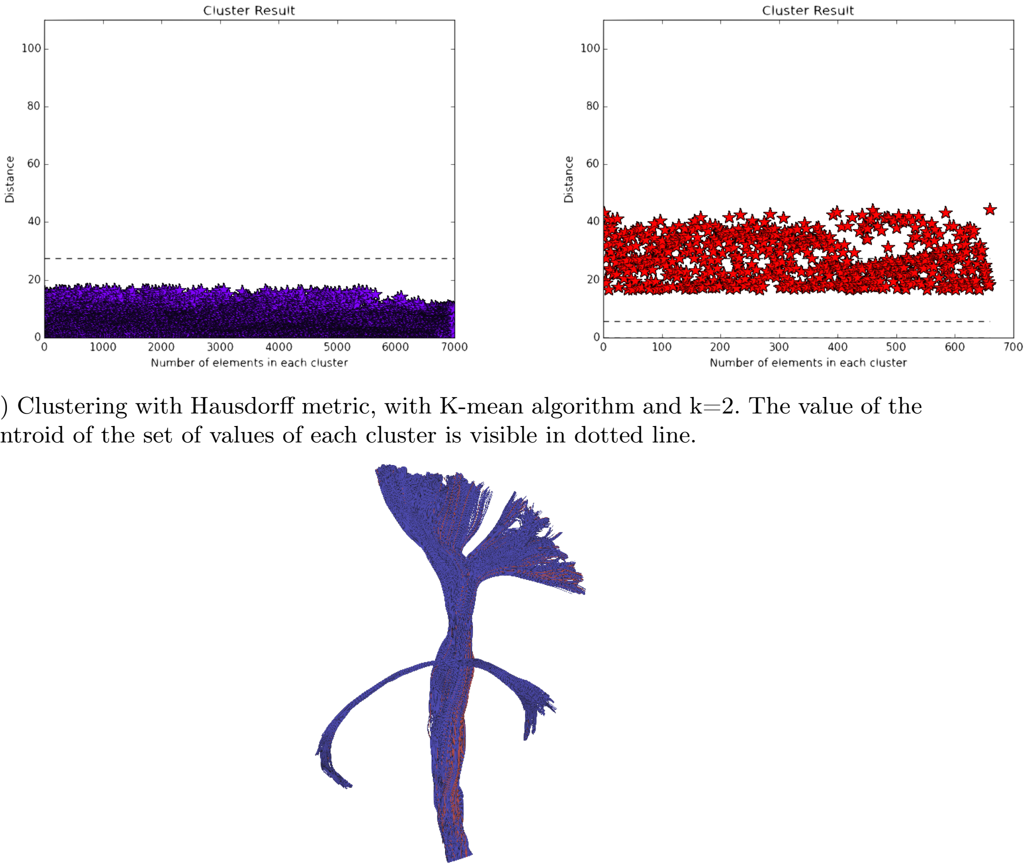


Fig. 5.16: CST and CA Dataset.

Figure(5.17a) shows the index of distance using Hausdorff metrics and K-mean algorithm and k=2; in violet color the result of the first cluster and in red color the result of the second. In x-axis we have the number of elements (fibers) and in y-axis we have the value of distance. The value of the centroid of the set of values of each cluster is visible in dotted line. Figure(5.17b) shows that the Hausdorff metric does not respect the classification of the fibers. In fact the CST bundle and the CA bundle are not distinctly recognized.

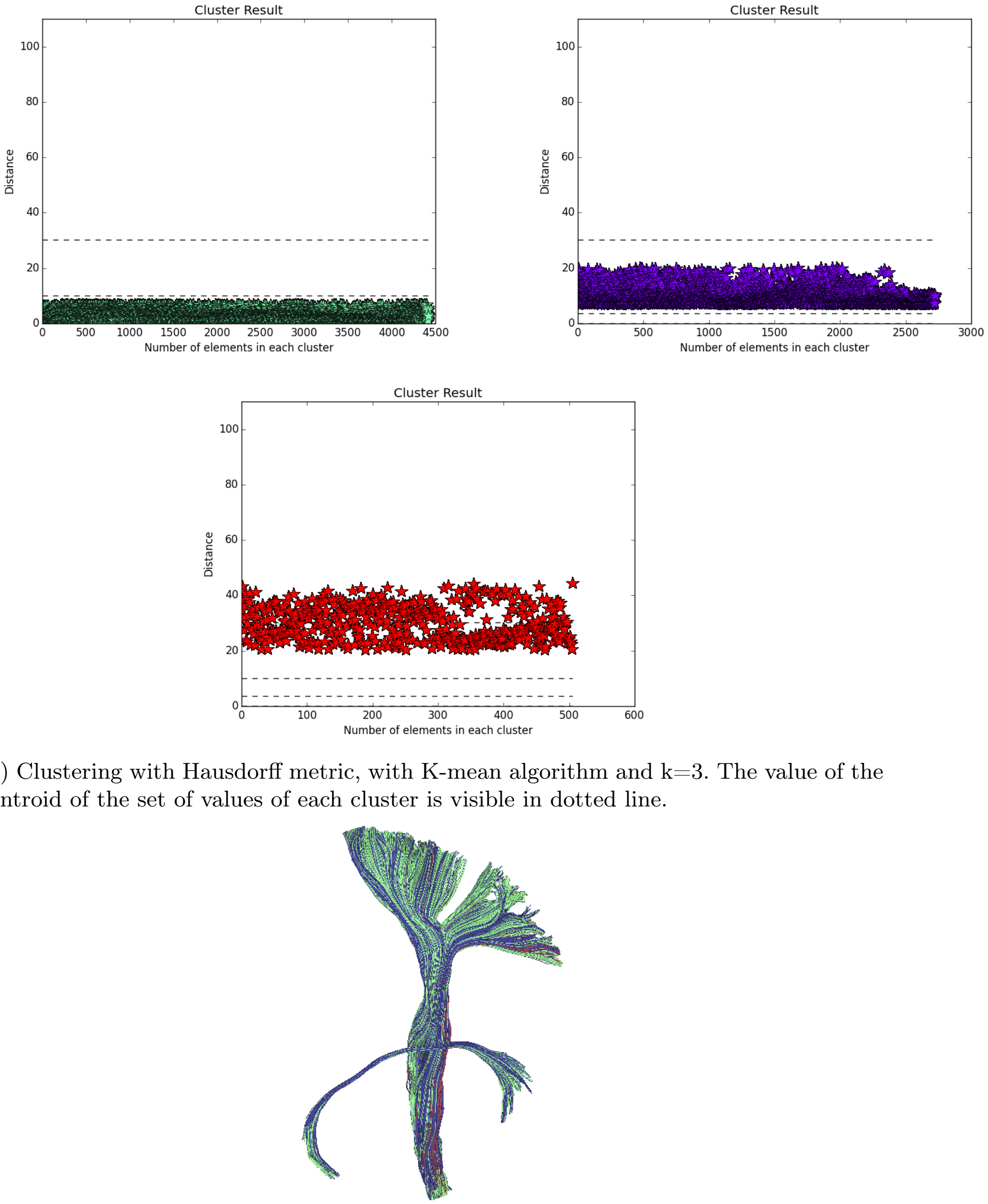


(a) Clustering with Hausdorff metric, with K-mean algorithm and k=2. The value of the centroid of the set of values of each cluster is visible in dotted line.

(b) CST and CA Fibers Clustering with Hausdorff metric.

Fig. 5.17: CST and CA Fibers Clustering with Hausdorff metric, and K-means algorithm (k=2).

Figure(5.18a) shows the index of distance using Hausdorff metrics and K-mean algorithm and k=3; in green color the result of the first cluster, in violet color the result of the second cluster and in red color the result of the third cluster. In x-axis we have the number of elements (fibers) and in y-axis we have the value of distance. Figure(5.18b) shows that the Hausdorff metric does not respect the classification of the fibers.

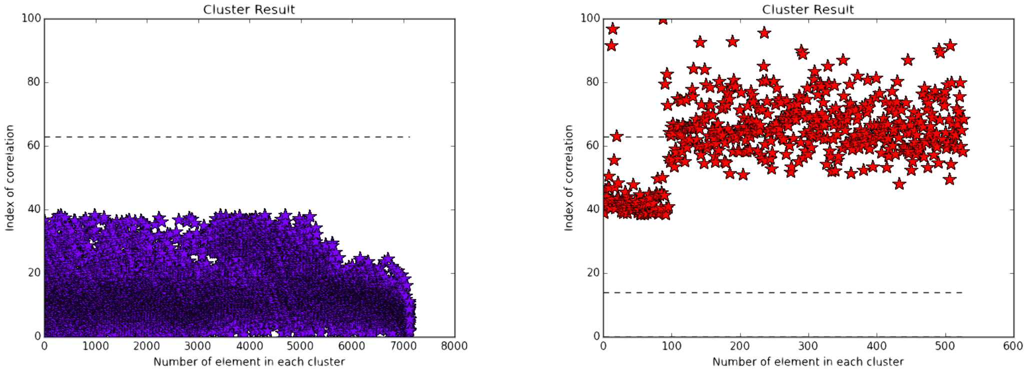


(a) Clustering with Hausdorff metric, with K-mean algorithm and k=3. The value of the centroid of the set of values of each cluster is visible in dotted line.

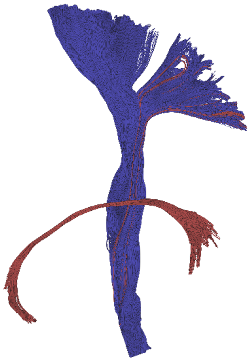
(b) CST and CA Fibers Clustering with Hausdorff metric.

Fig. 5.18: CST and CA Fibers Clustering with Hausdorff metric, and K-means algorithm (k=3).

Figure(5.19a) shows the index of correlation using NewSimilarity metrics and K-mean algorithm and k=2; in violet color the result of the first cluster and in red color the result of the second cluster. In x-axis we have the number of elements (fibers) and in y-axis we have the index of correlation. Figure(5.19b) shows a good classification of the fibers. As can we see in Figure(5.19a) (right image) few values of fibers in red color, are considered in the wrong bundle.



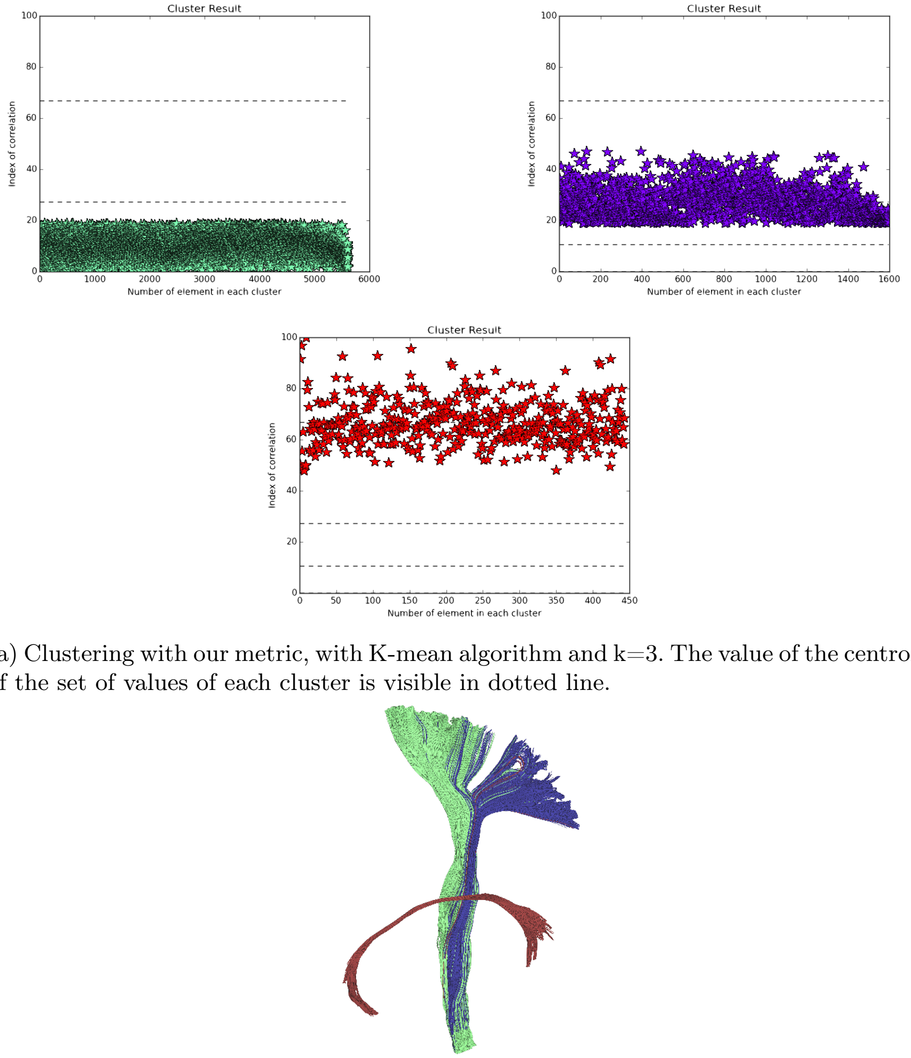
(a) Clustering with NewSimilarity metric, with K-mean algorithm and k=2.



(b) CST and CA Fibers Clustering with NewSimilarity metric.

Fig. 5.19: CST and CA Fibers Clustering with NewSimilarity metric, and K-means algorithm (k=2).

Figure(5.20a) shows the index of correlation using NewSimilarity metrics and K-mean algorithm and k=3; in green color the result of the first cluster, in violet color the result of the second cluster and in red color the result of the third cluster. In x-axis we have the number of elements (fibers) and in y-axis we have the index of correlation. The value of the centroid of the set of values of each cluster is visible in dotted line. Figure(5.20b) shows that the NewSimilarity metric, with value of parameter k equal three, classify very well CA bundle (red color) and CST bundle is subdivision in two sides.



(a) Clustering with our metric, with K-mean algorithm and k=3. The value of the centroid of the set of values of each cluster is visible in dotted line.

(b) CST and CA Fibers Clustering with NewSimilarity metric.

Fig. 5.20: CST and CA Fibers Clustering with NewSimilarity metric, and K-means algorithm (k=2).

Table(5.3), shows the tables that contain the index for evaluate the clustering performance. The indices we have considered are:

- homogeneity score(homo): each cluster contains only members of a single class.
- completeness score(compl): all members of a given class are assigned to the same cluster
- V-masure(v-meas): entropy-based measure which explicitly measures how successfully the criteria of homogeneity and completeness have been satisfied.
- adjusted Rand index(ARI): computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings.
- adjusted mutual information(AMI): function that measures the agreement of the two assignments, ignoring permutations.
- silhouette coefficient(silhouette): measures how close a fiber is to its own cluster in comparison to the rest of the clusters, i.e. whether there is another cluster that might represent it better or as well. Silhouette analysis can be used to study the separation distance between the resulting clusters.

The algorithm used for fibers classification is K-mean clustering, with k=2 and k=3; metrics used to analyze this dataset are Hausdorff metric and NewSimilarity metric.

Metric: Hausdorff
Clustering Algorithm: K-means

Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	70528	0.078	0.048	0.060	-0.088	0.048	0.831
n_cluster 3	34999	0.085	0.028	0.042	-0.037	0.027	0.643

(a) Clustering Performance with Hausdorff metric.

Metric: NewSimilarity
Clustering Algorithm: K-means

Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	46	0.048	0.023	0.032	0.018	0.023	0.646
n_cluster 3	22	0.239	0.074	0.113	0.004	0.074	0.610

(b) Clustering Performance with NewSimilarity metric.

Table 5.3: Clustering Performance using different metrics and K-means algorithm.

5.3 Conclusions

In this chapter we have reported the results obtained by applying the distance metrics to the datasets that have been selected come from data available online by ISMRM 2015 Tractography challenge - Data, in Figure (5.1), Figures (5.2) and Figures (5.3).

To each dataset the distance metrics of Frechet, Hausdammff, MDF and NewSimilarity were applied using two different clustering algorithms, K-means and Agglomerative clustering. The results obtained were subsequently evaluated through the clustering indices explained in the previous chapters.

From the results obtained on the data available online it come out that in different cases, using a metric that takes into account not only the distance between fiber pairs, but also their shape as elaborated in our "NewSimilarity Metric" gives back the best results in the clustering algorithms. The study of the shape parameter that we have performed shows how important it is, to have a good similarity metric, not to consider only the local component but also the global one. Indeed, two fibers are usually considered similar if they are separated by a small distance, have comparable length and similar shape. Motivated by this last point of view, in our approach, we use the combination between the shape and distance similarity highlighting the concept of shape among the fibers for clustering.

6

Algorithm verification with clinical data

6.1 Introduction: Clinical Context

To date, Diffusion Tractography (DT) remains the only non-invasive method for visualizing human brain connections. DT suffers from both fundamental and practical limitations that limit its use for modeling brain connections. Unlike many invasive modalities, DT is incapable of determining the direction of information flow, nor can it distinguish single-neuron and multi-neuron connections. DT may also have difficulty resolving complex intra-voxel fiber crossings or non-dominant fiber populations due to limitations in scan time, hardware, or processing methods. Despite its many limitations, DT has been successfully used to model human neuronal connections for over two decades, including several pathways that are putative Deep Brain Stimulation (DBS) targets.

There are many different DT data processing methods available, each with different requirements, assumptions, limitations and benefits. Several tractography algorithms have been described. In brief, commonly used methods are deterministic and probabilistic. Deterministic tractography uses a linear propagation approach that takes into account the main eigenvector direction of each voxel [5]. Among its main caveats is the potential misrepresentation of fibers in regions of crossing projections. Probabilistic fiber tracking involves seeding the same region numerous times so that the probability of a tract passing through each voxel may be calculated [6,7].

Diffusion Tensor Imaging (DTI) is by far the most common DT method used in DBS studies [66]. This technique has the advantage of being readily available on most commercial scanners, as well as having a relatively quick acquisition, and a simple reconstruction.

Deep brain stimulation is a well-studied therapeutic option for movement disorders such as Parkinson's disease, tremor, and dystonia. In addition, DBS is being increasingly investigated for many other disorders affecting the central nervous system.

In the clinic, tractography has been used in patients with tremor undergoing ventral intermediate nucleus (ViM) thalamic DBS. The VIM is not directly visualized on 3T images [12] and targeting relies on a combination of indirect measurements derived from different approaches and commonly used stereo-tactic atlases [13].

Figure(6.1) shows the workflow we have applied to the fiber similarity metrics for the classification of the anatomical bundles, to process clinical data. We consider the data after the bundle fibers have been reconstructed to which we apply the similarity metrics (distance metrics) and we study the DBS. This framework shows a possible clinical application of distance metrics. With Diffusion Tractography it is possible to segment the bundles of fibers and to identify the areas for Deep Brain Stimulation; as they are visible from the magnetic resonance image. Once the bundles have been identified and the region derived from the crossing of the fibers is identified, it is possible to study various pathologies such as Parkinson's disease.

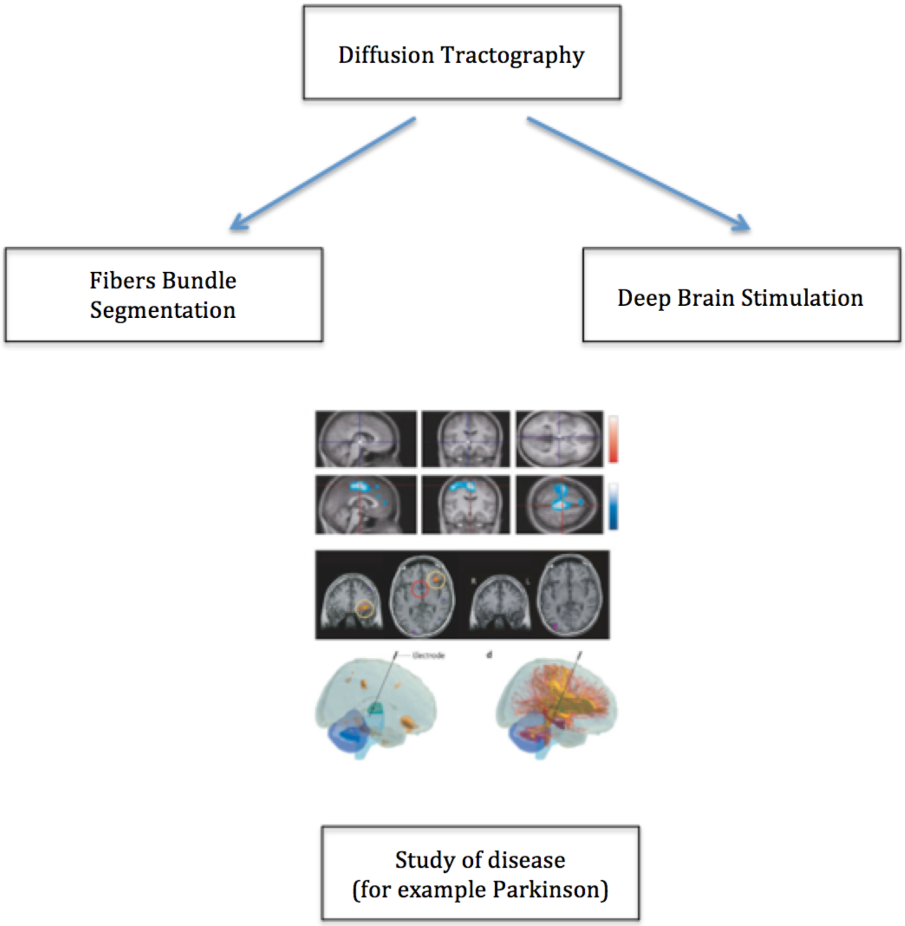

Diffusion Tractography
Fibers Bundle Segmentation
Deep Brain Stimulation
Electrode
d
Study of disease (for example Parkinson)

Fig. 6.1: Framework for a possible clinical application of distance metrics.

In particular, we see in detail in which step of the clinical pipeline the analysis of distance metrics is inserted. Before being able to apply the metrics described in the literature, for the classification of the anatomical bundles of the brain, a preliminary phase is performed. In this preliminary phase we found:

- acquisition of the dataset to be analyzed using the diffusion magnetic resonance
- use of different software for image pre-processing:
  - reduction of artifacts
  - choice of the combination of parameters (deterministic or probabilistic tractography which fiber reconstruction algorithm must be used) to have a clean dataset
  - manual extraction of the regions of interest to extract the anatomical bundles
- study of optimal software parameters:
  the neurosurgeon evaluates the result of the extraction of the fibers of the previous step based on his anatomical knowledge

Subsequently, the fibers of each bundles can be analyzed with the different distance metrics. This is the central point of the thesis.

- fibers bundle segmentation:
  fiber evaluation through the use of distance metrics. In this way, after having extracted the fibers with the manual method (user positioned ROI) the results are refined through the mathematical analysis which provides measurements of proximity of the area to be treated Deep Brain Stimulation; treatment may consist of the intersection of the anatomical bundles, of to reduce the tremor.

In the following we show the results obtained by analyzing, as a first step, only the step related to the similarity metric. The clinical data used have been previously processed using the tractography algorithms and the bundles have already been segmented.

The purpose is to understand which is the distance metrics that offers, compared to a ground truth, the best recognition of the different anatomical bundles. We also want to see if the metrics applied to online data are also applicable to clinical data.

Afterwards we will summarize in a table the results obtained by applying the selected metrics to the clinical datasets and how clustering occurs in the different cases. We will present in detail the most significant cases. We use the library in Python, "scikit-learn" for clustering classification.

6.2 Input Datasets

Figure(6.2) shows the bundles of clinical cases that we have considered for our analysis. In the first case we consider the Cortical Spinal bundle (left and right); in the second case we have the Cortical Spinal bundle (left and right) and the left Arcuate bundle and in the third case we have the Cortical Spinal (right) and Inferior fronto occipital fasciculus (right) bundles.

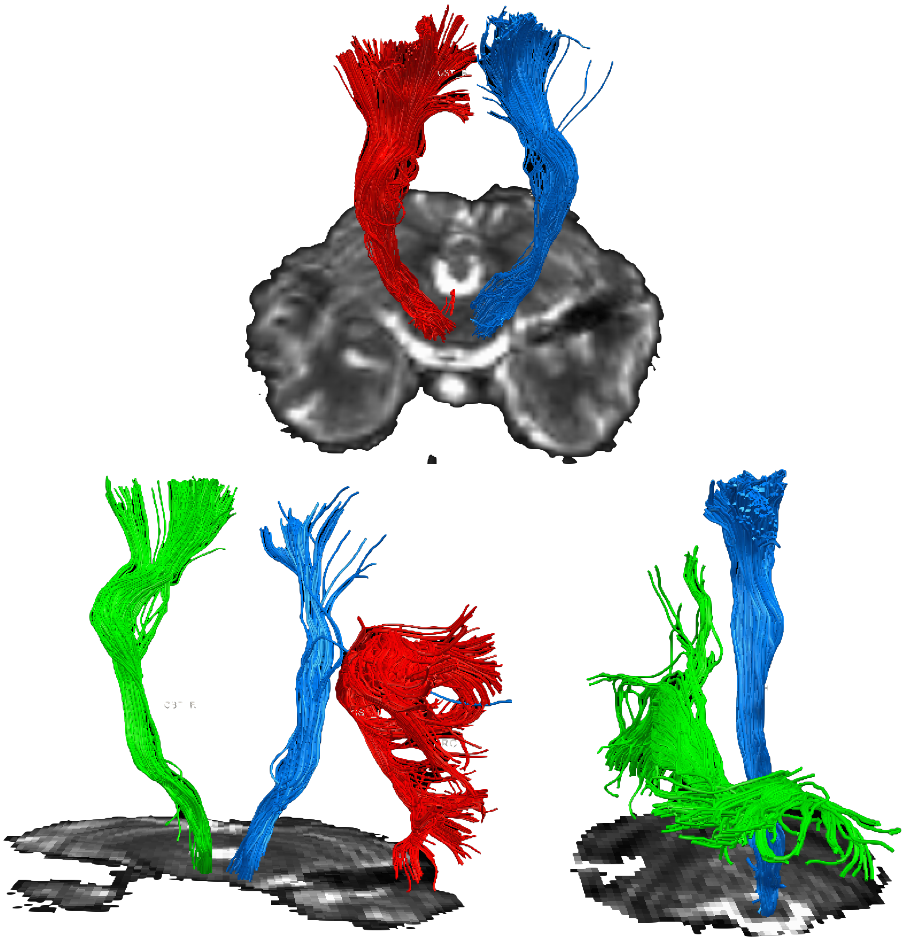


Fig. 6.2: At the top of the figure, the clinical case concerns the CST (left and right); in the left figure is represented the CST (left and right) and the Arcuate left; otherwise in the right figure is represented the CST and IFOF left.

The data were visualized with the TrackVis software and this allowed us to know the number of fibers of each bundle and the possibility of using this parameter as ground truth (GT) for the clustering algorithms. Regarding data processing, we have applied the same steps applied to the online data.

In the analysis phase, we applied to the selected fibers our new metrics and the other distance metrics most used in the literature. The results obtained through the clustering algorithms have been qualitatively evaluated by means of the quality indices defined in the previous chapter as we will see later summarized in the table. The indices used need as input of the algorithm a ground truth of the fibers i.e. the labels that identify the real position of the fibers in the anatomical bundle. The ground truths we used were those of dividing the labels according to the bundles considered. We want to identify the area, of the intersection of the anatomical bundles, to be treated to reduce the tremor. For this scope is not necessary to have an intra-bundles precision. In fact, from the medical point of view, there may be fibers that are not classified in the correct way for some errors in the data acquisition phase or errors due to the pre-processing phase. These errors do not compromise the clinical analysis.

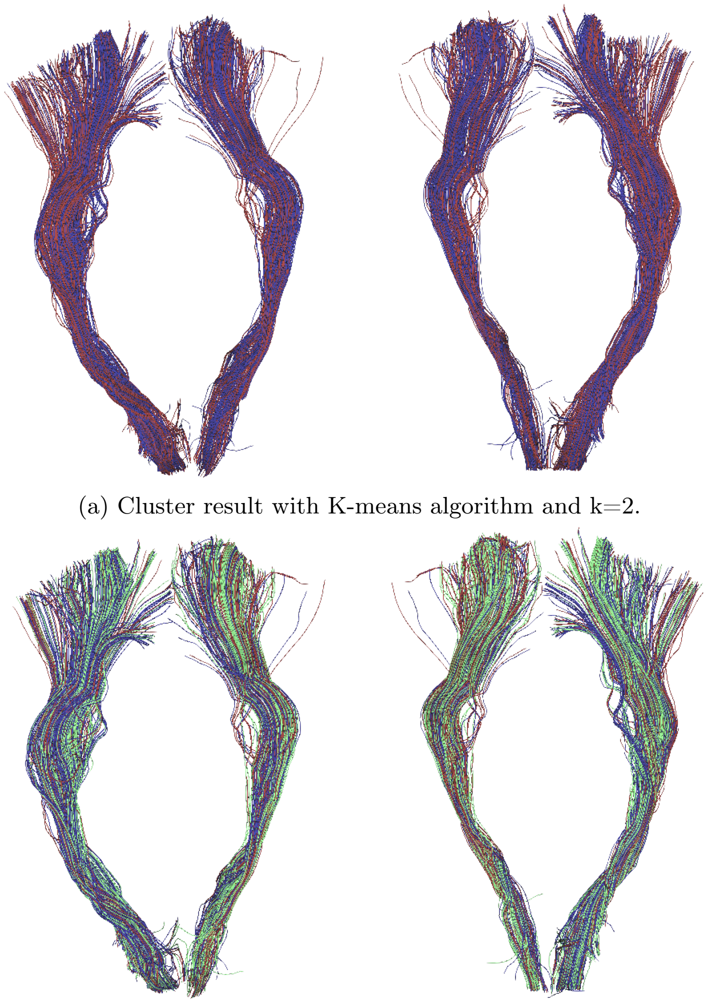
The important phase, and the one on which we have focused, is to apply the distance metrics seen in the previous chapter to identify the anatomical bundles by means of the distances between the fibers and to facilitate the identification of the underlying area of interest.

6.3 Results

Figure(6.3) shows the results of Table(6.1a), and Figure(6.4) shows results show in Table(6.1b). We have reported only the metrics that present significant indices of value and that allow to identify well the differences in the selection of the bundles. We analyze clustering performance and evaluate the indices the table shows best values when we use the NewSimilarity metric compared to the Frechet metric. This result is shown in Figure(6.3) and Figure(6.4b) where in the first case with Frechet metric, the right bundle is not distinguished from the left. The same result also occurs in the case, with the Frechet metric, in which the K-means algorithm is used with k = 3. In the second case, using the NewSimilarity metric, with either K-means algorithm with k=2 and K-means algorithm with k=3 the result of the classification of the bundles is more precise.

We consider the result of the NewSimilarity metric more precise because, as we can see in the figure, with our metric we can identify the right and left part of the bundle. If we consider the K-mans clustering algorithm with k = 2; in Figure(6.4b), there are some red color fibers incorporated in the violet color bundle, but according to the opinion of the neurosurgeon it is an error due to the pre-processing phase that does not depend on the metric type used and it does not have a clinical impact.

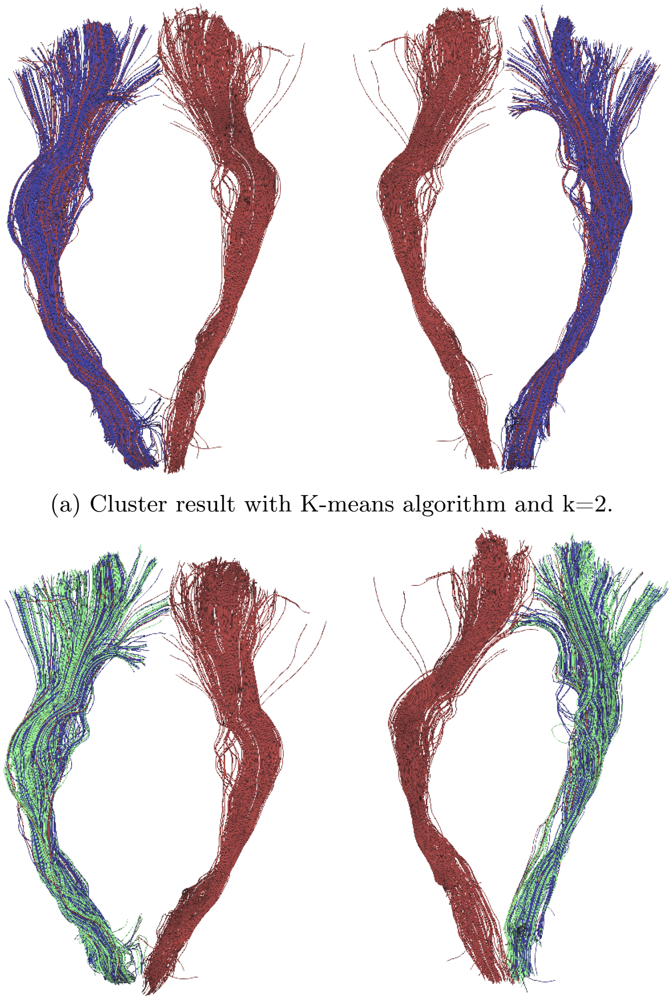
In the following we shows the results obtained with the K-means clustering algorithms using the Frechet metric and the NewSimilarity metric. In agreement with the neurosurgeon, with respect to the synthetic data, in the clinical data we report the obtained results Figure(6.3) and Figure(6.4) with anterior and posterior view of the beams. In this way there is a better view of the direction of the fibers.



(a) Cluster result with K-means algorithm and k=2.

(b) Cluster result with K-means algorithm and k=3.

Fig. 6.3: Cluster result with K-means algorithm and Frechet metric.



(a) Cluster result with K-means algorithm and k=2.

(b) Cluster result with K-means algorithm and k=3.

Fig. 6.4: Cluster result with K-means algorithm and NewSimilarity metric.

Table(6.1), shows the index to evaluate the clustering performance. The algorithm used for fibers classification is K-mean clustering and the analyzed metrics are Frechet metric and NewSimilarity metrics because with this metrics we have relevant results. Considering the data in the Table(6.1), NewSimilarity metric behaves better than Frechet metric. If we consider the indices of cluster performance, the results of the cluster algorithm, have good values and this can be seen in the fibers classification.

Metric: Frechet
Clustering Algorithm: K-means

Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	24393	0	0	0	-0.002	-0.001	0.618
n_cluster 3	11135	0.036	0.023	0.028	0.030	0.022	0.583

(a) Clustering Performance with Hausdorff metric.

Metric: NewSimilarity
Clustering Algorithm: K-means

Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	9	0.524	0.511	0.517	0.493	0.511	0.684
n_cluster 3	3	0.896	0.558	0.688	0.602	0.557	0.666

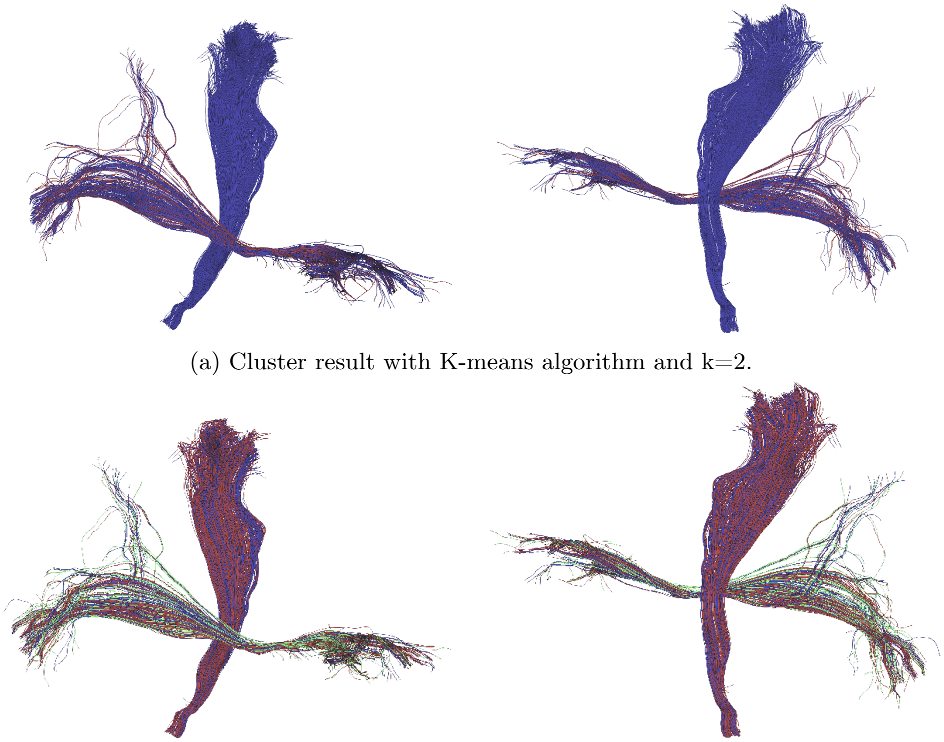
(b) Clustering Performance with NewSimilarity metric.

Table 6.1: Clustering Performance using different metrics and K-means algorithm.

Figure(6.5) shows the results of Table(6.2a), Figure(6.6) shows the results of Table(6.2b), Figure(6.7) shows the results of Table(6.2c) and Figure(6.8) shows the results of Table(6.2d).

Compared to the previous dataset, in this analysis we considered four similarity metrics, as described above. This choice is due to the greater difficulty in classifying different bundles, since the initial dataset shown in Figure (2) is composed of anatomical bundles that are spatially close to each other. Consequently, in this type of dataset, compared to that in Figure (1) classification is more problematic but of routine in clinical application. We are therefore interested in seeing how our metric behaves in the presence of a real clinical case like this.

Figure(6.5) shows the results obtained with K-means algorithm with Frechet metric. Using clustering algorithm with k = 2, the metric tends to identify both bundles with the same color; there are few red fibers that identify the horizontal bundle. Even in case of clustering algorithm with k = 3 the metric seems to behave well. The splitting of the dataset into three bundles appears disordered. With k = 2, the few red fibers are not considered important errors due to the pre-processing phase of the data, as previously mentioned; this we can affirm it because the metric tends to group the bundles with the same color, in this case violet color.

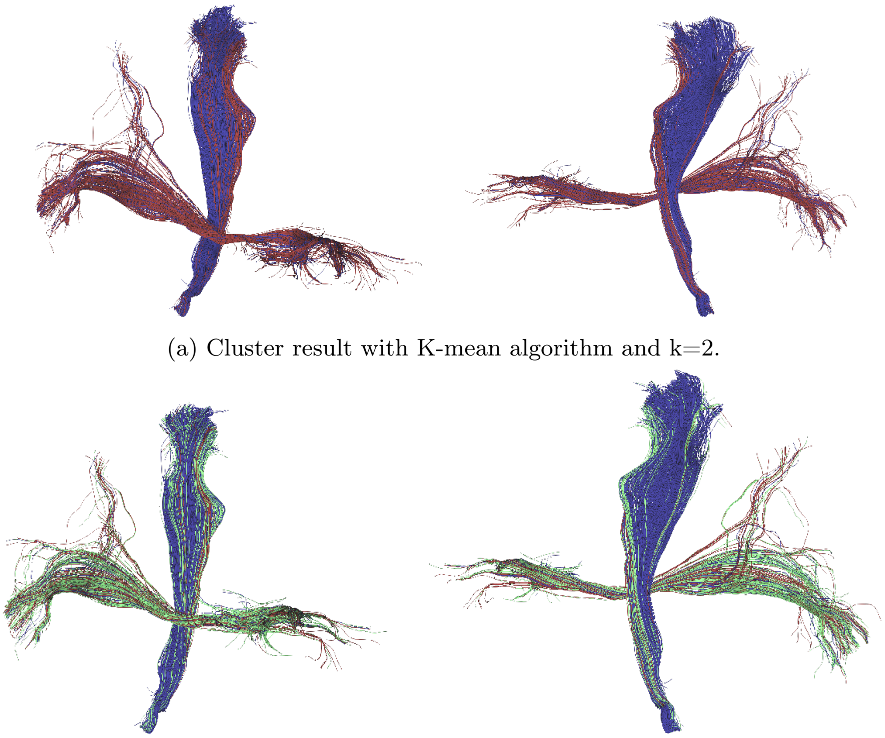


(a) Cluster result with K-means algorithm and k=2.

(b) Cluster result with K-means algorithm and k=3.

Fig. 6.5: Cluster result with K-mean algorithm and Frechet metric.

Figure(6.6) shows the results obtained with K-means algorithm with Hausdorff metric. In this case, clustering algorithm with k = 2, the metric tends to identify with red color the horizontal bundle and violet color the vertical bundle. Some red fibers are inserted inside the violet colored bundle, but this result compared to the Frechet metric is better because the metric identifies, even if not completely correct, the two anatomical bundles. In the case of clustering algorithm with k = 3 we have no significant improvement over the results obtained with the Frechet metric.

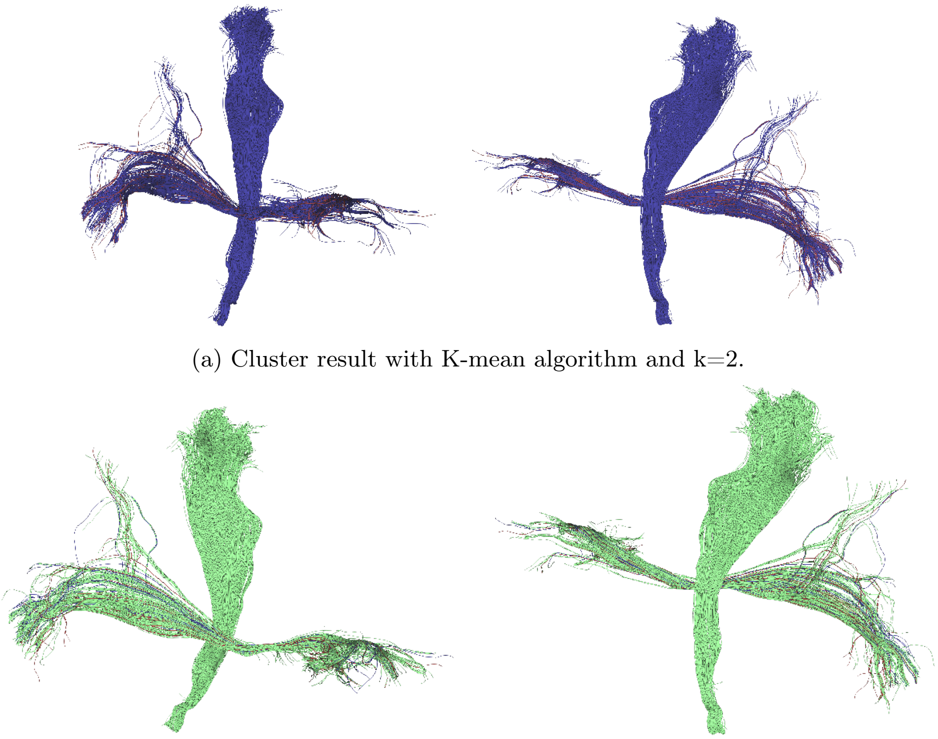


(a) Cluster result with K-mean algorithm and k=2.

(b) Cluster result with K-mean algorithm and k=3.

Fig. 6.6: Cluster result with K-mean algorithm and Hausdorff metric.

Figure(6.7) shows the results obtained with K-means algorithm with LCSS metric. In this case with algorithm clustering with k = 2 and algorithm clustering with k = 3, LCSS metric does not produce good results. We have introduced this type of metric because the LCSS metric, as described in chapter two, gives more importance to the concept of form. The Frechet and Hausdorff metrics used in the previous dataset and online datasets give more importance to the concept of point-to-point distance.

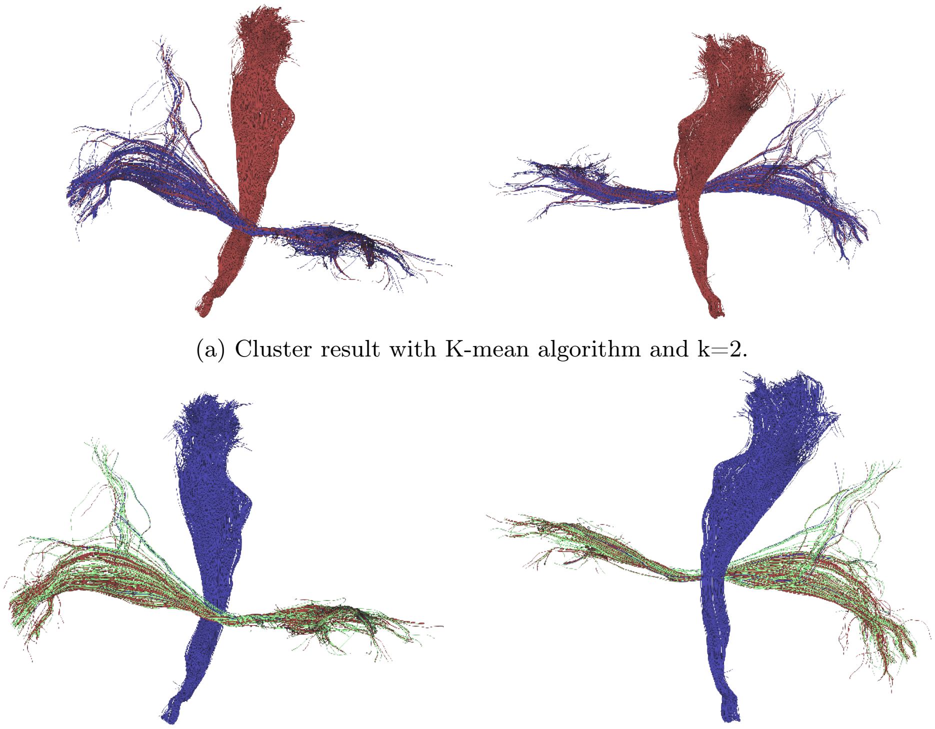


(a) Cluster result with K-mean algorithm and k=2.

(b) Cluster result with K-mean algorithm and k=3.

Fig. 6.7: Cluster result with K-mean algorithm and LCSS metric.

Figure(6.8) shows of results obtained with K-means algorithm with NewSimilarity metric. In the case of clustering algorithm with k = 2, the metric identifies the vertical bundle with the red color and the horizontal bundle with the violet color. Also in the case of k = 3 we have, compared to the previous metrics the best results. We obtain better results because if we compare the results obtained with algorithm clustering with k = 3 compared to the other metrics, we see a better classification of the fibers. In fact, for example with the NewSimilarity metric the vertical and horizontal bundles are classified more precisely than the previous results.



(a) Cluster result with K-mean algorithm and k=2.

(b) Cluster result with K-mean algorithm and k=3.

Fig. 6.8: Cluster result with K-mean algorithm and NewSimilarity metric.

Table(6.2), shows the tables that contain the index for evaluate the clustering performance. The algorithm used for fibers classification is K-mean clustering and the analyzed metrics are Frechet metric, Hausdorff metric, LCSS metric and NewSimilarity metric because with this metrics we have relevant results. Considering the data in the Table(6.2), NewSimilarity metric behaves better than others metrics; the results of the cluster algorithm, if we consider the indices of cluster performance, have good values and this can be seen in the different fibers classification.

Metric: Frechet
Clustering Algorithm: K-means

Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	241910	0.139	0.250	0.179	0.271	0.138	0.817
n_cluster 3	59604	0.139	0.084	0.105	0.115	0.083	0.714

(a) Clustering Performance with Frechet metric.

Metric: Hausdorff
Clustering Algorithm: K-means

Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	17677	0.215	0.196	0.205	0.333	0.195	0.688
n_cluster 3	9411	0.264	0.168	0.205	0.339	0.167	0.648

(b) Clustering Performance with Hausdorff metric.

Metric: LCSS
Clustering Algorithm: K-means

Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	0	0.063	0.165	0.091	o.147	0.062	0.963
n_cluster 3	0	0.077	0.148	0.101	0.173	0.075	0.968

(c) Clustering Performance with LCSS metric.

Metric: NewSimilarity
Clustering Algorithm: K-means

Cluster	Inertia	Homogeneity	Completeness	V-meas	ARI	AMI	Silhouette
n_cluster 2	9	0.452	0.478	0.465	0.631	0.452	0.815
n_cluster 3	3	0.675	0.78	0.560	0.744	0.478	0.804

(d) Clustering Performance with NewSimilarity metric.

Table 6.2: Clustering Performance using different metrics and K-means algorithm.

6.4 Conclusions

In this section we have described the results obtained by applying some of the similarity metrics for brain fiber classification of the analysis of clinical datasets. The results obtained with the NewSimilarity metric are good in terms of quality as can be seen in the images where the fibers are colored according to the clusters, and by evaluating the performance indices of the cluster. Figure(6.8) shows the results obtained with K-means algorithm with NewSimilarity metric. These results outperform the other metrics available in the literature and confirm the applicability of the new method to clinical cases. With the use of NewSimilarity metric for fibers classification we obtain results that agree with the anatomy of the brain bundles.

7

Conclusions and future work

7.1 Introduction

In this thesis, the field of application of our analysis was introduced in the first chapter. To recognize the anatomic bundles within the brain is a very important task in the scientific community. It is important both in the diagnostic phase, because a good display of the subdivision of the areas of the brain may help the surgeon in the general view of the problem and in the intra-operative phase since the identification the fascicles is the basis for a good result in neurosurgical resection of the tumor. From MR data, the white matter fiber tract can be reconstructed using a class of technique called tractography. Each dataset derived by tractography is composed of a large number of streamlines, which are sequences of points in 3D space. To simplify the visualization and analysis of white matter fiber tracts obtained from MR data, it is often necessary to group them into larger clusters or bundles. In order to perform clustering, first a mathematical definition of fiber similarity (or more commonly a fiber distance) must be specified. Then, pairwise fiber distance may be calculated and used as input for a clustering algorithm. We have defined a new metric called "NewSimilarity" to assess the similarity of tracts, we have established the right number of clusters and we have evaluated, once the clusters are obtained, qualitatively and quantitatively the results. The final scope of this analysis is to obtain a new metric for brain fiber classification. In subsequent chapters, we analyze the similarity metrics among fibers best known in the literature and we have studied the main clustering methods. Then, we summarize the mathematics used to describe our method, the obtained results, and we explain the possible field of application of the NewSimilarity metric.

7.1.1 Mathematical Background

After a brief introduction explaining metric spaces and the proprieties needed to talk about distance, we have introduced the concept of curve, for planar and three dimensional cases. Considering the fibers of the brain as a set of points that form generic curves, we analyze the behavior of the curves in 3D space; then, we introduced the concept of correlation and its proprieties. Specifically we described cross-correlation, which is the function used to evaluate our similarity metric.

The most common measures used for distance only capture the local relationship between streamlines but not the global structure of the fiber. Global structure, refer to the fiber variability shape. Together, local and global information, may define a good measure of similarity.

Therefore, to analyze the shape of fibers, we introduced two parameters, linear distance between points and angular shape, that describe the new geometric distance computed for each pair of fibers. The values of the tangents are obtained from the Frenet frame previously calculated. In order to obtain even spatial information on the fiber pattern, we computed the internal product of the tangents values. The distance between the fibers is calculated from the standard Euclidean definition. The point coordinates refer to the reference frame of the data acquisition system, in our case to the magnetic resonance system. For each fiber in the bundle the similarity is calculated for each pairs of points in two different fibers. Two fibers belong to the same bundle, if the distance separating them is small if they have similar shapes. The criterion used to classify fibers is based on the correlation concept.

7.1.2 Results

For our analysis we have considered two types of datasets; the first dataset is available online by ISMRM 2015 Tractography challenge - Data. The second is a clinical dataset. We have considered for analysis only some of the metric described in the previous chapters; Frechet metric, Hausdorff metric, MDF metric and our metric that we call "NewSimilarity" metric. We have done a quantitative analysis of the dataset. The similarity metrics are applied to the initial dataset and a clustering algorithm is chosen for the classification of the fibers. We use two graphs to represent this analysis; the first one where the fibers of the initial dataset are colored after the clustering. In this way we have a first visual result of the classification of the fibers. The second graph, where in x-axis we represent the value of similarity fibers and in y-axis we represent the number of element(fibers) in each cluster: by comparing the similarity values, we see the corresponding number of fibers classified within the cluster. This second analysis that we have done is quantitative and to facilitate the comparison of the different metrics, we report a summary table of the clustering performance with the index values. The results confirm that by considering distance and shape informations in fibers brain classification we obtain better results than by using distance alone. In fact, this is the key of where, in addition to the linear distance between points, we add the "Angular Shape" component.

7.1.3 Future Work

To date, Diffusion Tractography (DT) remains the only non-invasive method for visualizing human brain connections. DT suffers from both fundamental and practical short comings that limit its use for modeling brain connections. Unlike many invasive modalities, DT is incapable of determining the direction of information flow, nor can it distinguish single-neuron and multi-neuron connections. DT may also have difficulty resolving complex intra-voxel fiber crossings or non-dominant fiber populations due to limitations in scan time, hardware, or processing methods. Despite these limitations, DT has been successfully used to model human neuronal connections for over two decades, including several pathways that are putative Deep Brain Stimulation (DBS) targets. Diffusion Tensor Imaging (DTI) is by far the most common DT method used in DBS studies.

We have seen where the analysis of distance metrics can be inserted in the clinical pipeline. Before applying the metrics described in the literature, to classify the anatomical bundles of the brain, a preliminary test was performed: the acquisition of the dataset using the diffusion magnetic resonance, use of different software for image pre-processing, study of optimal software parameters and fibers bundle classification. Fibers of each bundles can be analyzed with the different distance metrics and this is the main contribution of the thesis as a future work. We want to study the phases prior to the fiber segmentation. In fact, in order to identify the area to perform DBS to reduce the tremor in some pathologies, it is necessary to understand which are the best data acquisition and pre-processing parameters and which facilitate the segmentation of the anatomical bundles. This task requires first to apply the deterministic and probabilistic tract algorithms to the data acquired through the DTI; second, on the basis of these algorithms, we need to understand which work pipeline is best for the similarity metrics applied.

After applying the similarity metrics to the fibers, that make up the anatomical bundles of the brain, the volume occupied by these bundles must be identified: the intersection volume of the bundles gives us information about the area to be treated with DBS. A good estimate of distance between the anatomical bundles is useful to localize the area for stimulation and identify the boundaries of the bundles. It is necessary to know the distance between the inner fibers, and in this thesis we have analyzed this phase. We hope that in a near future we will be able to apply this approach to real clinical cases.

Sommario

Le tecniche di diffusione di immagini a risonanza magnetica (dMRI) forniscono un modo non invasivo per esplorare l’organizzazione e l’integrità delle strutture della materia bianca nel cervello umano. dMRI quantifica in ciascun voxel, il processo di diffusione delle molecole d’acqua che sono costrette meccanicamente nel loro movimento dagli assoni dei neuroni. Questa tecnica può essere utilizzata nella pianificazione chirurgica e nello studio della connettività anatomica, dei cambiamenti cerebrali e dei disturbi mentali. Dai dati di dMRI, i tratti di fibra della materia bianca possono essere ricostruiti usando una classe di tecnica chiamata tractography. Il set di dati derivato dalla trattografia è composto da un gran numero di fibre o streamlines, che sono sequenze di punti nello spazio 3D. Per semplificare la visualizzazione e l’analisi di tratti di fibre di materia bianca ottenuti con algoritmi di trattografia, è spesso necessario raggrupparli in cluster o bundles anatomici. Questo passaggio è chiamato clustering. Per eseguire il clustering, è necessario fornire una definizione matematica della similarità delle fibre (o più comunemente una distanza tra le fibre). Sulla base di questa metrica, la distanza tra coppie di fibre può essere calcolata e utilizzata come input per gli algoritmi di clustering. Le metriche più comuni utilizzate per la misurazione della distanza sono in grado di acquisire solo la relazione locale tra le fibre del cercello ma non la struttura globale della fibra. La struttura globale si riferisce alla variabilità della forma. Insieme, le informazioni locali e globali, possono definire una migliore metrica di somiglianza. Abbiamo estratto le informazioni globali utilizzando una rappresentazione matematica basata sullo studio del tratto con equazioni di Fr ’enet. In particolare, abbiamo definito alcuni parametri intrinseci delle fibre che hanno portato a una classificazione dei tratti basata su caratteristiche geometriche globali. Utilizzando questi parametri, è stata sviluppata una nuova metrica di distanza per la somiglianza delle fibre. Per la valutazione della bontà della nuova metrica, sono stati utilizzati degli indici per uno studio qualitativo dei risultati.

References

1. *http://www.humanconnectomeproject.org.*
2. *Convolution and Correlation.* Springer, 2010.
3. Einstein A. *Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen.* Ann Phys., 1905.
4. H. Park M. Shenton. A. Brun, H. Knutsson. Clustering fiber tracts using normalized cuts. pages 368–375. In MICCAI'04, Conf., Lecture Notes in Computer Science., 2004.
5. Hans Knutsson Anders Brun, Hae-Jeong Park and Carl-Fredrik Westin. Coloring of dt-mri fiber traces using laplacian eigenmaps. *In EUROCAST'03, Conf. Proc., Lecture Notes in Computer Science.*, Springer Verlag.(2809.):564–572, 2003.
6. Alfred O. Hero Alan S. Willsky. Andy Tsai, Carl-Fredrik Westin. Fiber tract clustering on manifolds with dual rooted-graphs. *IEEE Trans.*, 2007.
7. Fumagalli L et al. Artico M, Spoletini M. Egas moniz: 90 years (1927–2017) from cerebral angiography. *Fronties in Neuroanatomy*, 2017.
8. Anderson AW. Theoretical analysis of the effects of noise on diffusion tensor imaging. *Magn Reson Med*, pages 46(6):1174–1188, 2001.
9. Townsend P. Valk Bailey, D. L. Positron-emission tomography: Basic sciences. ISBN 1-85233-798, 2005.
10. Sharma R Gupta AK. Baliyan V, Das CJ. Diffusion weighted imaging: Technique and applications. world journal of radiology. *World Journal of Radiology.*, (8(9):785-798), 2016.
11. Lohkamp Joachim Schwarz Matthias Bär, Christian. *Global Differential Geometry.* Springer, 2012.
12. Mirela Ben-Chen. *Differential Geometry of Curves.*
13. Christopher M. Bishop. *Pattern Recognition and Machine Learning.* 2006.
14. Victor Bryant. *Metric Sapces: Iteration and Application.*
15. X. He S. Mai. C. Bohm, J. Feng. A novel similarity measure for fiber clustering using longest common subsequence. *KDD.*, 2011.
16. Pin Chang. C. R. Wilke. Correlation of diffusion coefficients in dilute solutions. 1., June 1955.
17. de Schotten T. Catani M. A diffusion tensor imaging tractography atlas for virtual in vivo dissections. 2008.
18. Thiebaut de Schotten M. Catani M. A diffusion tensor imaging tractography atlas for virtual in vivo dissections. *PubMed.*, 2002.
19. Lin-Ching Chang, Derek K. Jones, and Carlo Pierpaoli. Restore: Robust estimation of tensors by outlier rejection. *Magnetic Resonance in Medicine*, 53(5):1088–1095, 2005.

20. Yi-Ping Chao, Jyh-Horng Chen, Kuan-Hung Cho, Chun-Hung Yeh, Kun-Hsien Chou, and Ching-Po Lin. A multiple streamline approach to high angular resolution diffusion tractography. *Medical Engineering and Physics*, 30(8):989–996, Oct 2008.
21. Cetingul H. E. Demir A., Mohamed A. Online agglomerative hierarchical clustering of neural fiber tracts. In *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 85–88, 2013.
22. Anderson AW. Ding Z, Gore JC. Classification and quantification on neuronal fiber pathways using diffusion tensor mri. *Magnetic Resonance in Medicine.*, (716-721), 2003.
23. M. Correia G. Williams I. Smith. E. Garyfallidis, M. Brett. Quickbundles, a method for tractography simplification. *PMC.*, 2012.
24. Aaron Filler. The history, development and impact of computed imaging in neurological diagnosis and neurosurgery: Ct, mri, and dti. Nature Precedings., 12 July 2009.
25. Rivière D. Guevara P., Poupon C. Robust clustering of massive tractography datasets. *Neuroimage.*, 54:1975–1993., 2011.
26. R. Bolles H. Barrow, J. Tenenbaum and H. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. 1977.
27. P. Scheuermann X. Wang H. Ding, G. Trajcevski and E. Keogh. Querying and mining of time series data: Experimental comparison of representations and distance measures. *VLDB*, 1:1542–1552, 2008.
28. Peter F. Neher Frederik B. Laun Bram Stieltjes Klaus H. Maier Hein. Fiberfox: Facilitating the creation of realistic white matter software phantoms. *Magnetic Resonance in Medicine.*, 09 December 2013.
29. Jeroen Hendrikse. *CT and MRI Scans: The Basic Principles.* 2017.
30. Walter Huda and R. Brad Abrahams. Radiographic techniques, contrast, and noise in x-ray imaging radiographic techniques, contrast, and noise in x-ray imaging. *American Journal of Roentgenology.*, 204.(2.), 2015.
31. D. Huttenlocher and W. Rucklidge. A multi-resolution technique for comparing imags using the hausdorff distance. volume CVPR, 1993.
32. S. Gouttard I. Corouge and G. Gerig. Towards a shape model of white matter fiber bundles using diffusion tensor mri. pages 344–347, 2004.
33. P. Sellerc M. Vealec P. Sellinb S. Jacquesd J. Scuffhama, D. Wilsonc and J. Cernikd. A cdte detector for hyperspectral spect imaging. *Journal of Instrumentation.*, 30 August 2012.
34. Q. Yang C. Bohm A. M. Wohlschlager N. Myers J. Shao, K. Hahn and C. Plant. Combining time series similarity with density-based clustering to identify fiber bundles in the human brain. *In ICDM Workshops.*, pages 747–754., 2010.
35. Peter Kostelec Jeffrey B. Woodward Javed A. Aslam Daniela Rus Daniel Rockmore Michael S. Gazzaniga. John D. Van Horn, Jeffrey S. Grethe. The functional magnetic resonance imaging data center (fmridc): the challenges and rewards of large–scale databasing of neuroimaging studies. *The Royal Society.*, 356.:1323–1339., 29 August 2001.
36. Stein Jonathan. *Correlation, in Digital Signal Processing: A Computer Science Perspective.* 2000.
37. Ioannis Karatzas and Steven. Shreve. *Brownian motion and stochastic calculus.* Springer Science, Business Media., 2012.
38. Elster AD Kilgore Ej. Walter dandy and the history of ventriculography. *Radiology*, (194):657–60, 1995.
39. J. Roussel SA van Bruggen. King, MD. Houseman. *q-Space imaging of the brain.*, volume 32. Magn.Reson.Med., 1994.

40. D. Aquino. L. Minati. *Probing neural connectivity through Diffusion Tensor Imaging (DTI).* Cybernetics and Systems., 2006.
41. Paul C Lauterbur and Peter Mansfield. Magnetic resonance imaging. *All Nobel Prizes in Physiology or Medicine.*, 6 October 2003.
42. E. Breton. Le Bihan, D. *Imagerie de diffusion in-vivo par résonance magnétique nucléaire.*, volume 301. 1985.
43. E. Breton E. Lallemand D. Aubin ML. Vignaud. Le Bihan, D. *Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging.*, volume 168. Radiology, 1988.
44. E. Breton E. Lallemand D. Grenier P. Cabanis. Le Bihan, D. *MR imaging of intravoxel inchoerent motions: application to dissusion and perfusion in neurologic disorders.* Radiology., 1986.
45. Guo L. Liu T. Hunter J. Wong S. T. Li H., Xue Z. A hybrid approach to automatic clustering of white matter fibers. *Neuroimage.*, 49, 2010.
46. L. Grimson S. Warfield W. Wells. M. Maddah, W. Eric. A unified framework for clustering and quantitative analysis of white matter fiber tracts. *Medical Image Analysis.*, 12, Issue2.:191–202., 2008.
47. D. Gunopulos M. Vlachos and G. Kollios. Discovering similar multidimensional trajectories. *In ICDE.*, 2002.
48. K. H. et al.. Maier Hein. Tractography challenge ismrm 2015 data.
49. Shadmi R. Batikoff A. Greenspan H. Mayer A., Zimmerman-Moreno G. A supervised framework for the registration and segmentation of white matter fiber tracts. *IEEE Trans. Med. Imaging*, 30(131-145), 2011.
50. A. J. H. McGaughey and M. Kaviany. Quantitative validation of the boltzmann transport equation phonon thermal conductivity model under the single-mode relaxation time approximation. *Phys. Rev. B*, 69:12, Mar 2004.
51. Donald W. McRobbie. Mri from picture to proton. *New York: Cambridge University Press.*, ISBN 0-521-68384-X, 2007.
52. Bahram Ravani Michael G. Wagner. Curves with rational frenet-serret motion. *Elsevier*, 15:79–101, 1997.
53. van Wijk JJ. Moberts B, Vilanova A. Evaluation of fiber clustering methods for diffusion tensor imaging. IEEE, 2005.
54. Van Zijl P. Nagae-Poetscher L. Mori S., Wakana S. Mri atlas of human white matter. *Neuroradiol.*, 27(1384-1385), 2005.
55. Shenton M. E. O'Donnell L., Kubicki M. A method for clustering white matter fiber tracts. *Neuroradiol.*, 27.:1032–1036., 2006.
56. PhD Lisa Jonasson PhD Philippe Maeder MD Jean-Philippe Thiran PhD Van J. Wedeen MD Patric Hagmann, MD and PhD Reto Meuli, MD. Understanding diffusion mr imaging techniques: From scalar diffusion-weighted imaging to diffusion tensor imaging and beyond. *RadioGraphics*, 26, 2006.
57. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. scikit-learn scikit-learn. machine learning in python. *Journal of Machine Learning Research*, 12.(2825–2830.), 2011.
58. J. Philibert. *One and a half century of diffusion: Fick, Einstein, before and beyond.* Diffusion Fundamentals., 2005.
59. Stojanovic-Radic J Yue GH Pioro EP-et al. Rajagopalan V, Jiang Z. A basic introduction to diffusion tensor imaging mathematics and image processing steps. *Brain Disord Ther 6:229.*, 2017.
60. Stenzel M. Mentzel H.-J. Reichenbach J. R. Ros C., Güllmar D. Atlas-guided cluster analysis of large tractography datasets. 2013.

61. Annette Ruckstuhl. Thomas graham's study of the diffusion of gases. *J. Chem. Educ.*, 28 (11).(594.), 1951.
62. Micheal O Searcoid. *Metric Sapces.* 2007.
63. Theodore Shifrin. Differential geometry: A first course in curves and surfaces. 2016.
64. Medina S. Varoquaux G. Thirion B. Siless, V. A comparison of metrics and algorithms for fiber clustering. *PRNI*, pages 190–193., 2013.
65. Alfred Gray Simon Salamon, Elsa Abbena. *Modern Differential Geometry of Curves and Surfaces with Mathematica.* 1998.
66. Cristina V. Torresa Rafael Manzanaresb Rafael G. Solac. Integrating diffusion tensor imaging-based tractography into deep brain stimulation surgery: A review of the literature. *Stereotact Funct Neurosurg*, 92:282–290, 2014.
67. Claudia Plant. Son T Mai, Sebastian Goebl. A similarity model and segmentation algorithm for white matter fiber tracts. *IEEE.*, 2012.
68. Michael Spivak. *A comprehensive Introduction to DIFFERENTIAL GEOMETRY*, volume Volume one. 1999.
69. William Stillwell. *An Introduction to Biological Membranes: From Bilayers to Rafts.* 2013.
70. Loring W. Tu. *Differential Geometry: Connections, Curvature, and Characteristic Classes.* Springer, 2017.
71. Panzenboeck M. M. Fallon J. H. Perry M. Gollub R. L. et al. Wakana S., Caprihan A. Reproducibility of quantitative tractography methods applied to cerebral white matter. *Neuroimage.*, 36(630-644), 2007.
72. Partic; Tseng Wen-Yih Isaac; Reese Timothy. Wedden, Van; Hagmann. *Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging.*, volume 54. Magnetic Resonance in Medicine., 2005.
73. Dietmar Pfeifer Wolfgang Gaul. *From Data to Knowledge. Theoretical and Practical Aspects of Classification, Data Analysis and Knowledge Organization.* 1996.
74. Peter J. Basser Yaniv Assaf. Composite hindered and restricted model of diffusion (charmed) mr imaging of the human brain. *NeuroImage*, 27:48–58, 2005.
75. Laidlaw D. H. Zhang S., Correia S. Identifying white-matter fiber bundles in dti data using an automated proximity-based fiber-clustering method. *IEEE Trans. Vis. Comput. Graph.*, 14.:1044–1053, 2008.
76. Laidlaw DH. Zhang S, Demiralp C. Visualizing diffusion tensor mr images using streamtubes and stream-surfaces. *IEEE Trans. Visualization and Computer Graphics.*, pages 454–462, 2003 Oct.-Dec.