Jean Monnet University

Master Thesis

CPE Lyon

Projet de fin d'étude

Optimal transport for comparing short brain connectivity between individuals

---

Author
Duy Anh Philippe Pham

Under the supervisors of
Prof. Olivier Coulon
Academic Tutor
Prof. Marion Foare
Prof. Marc Sebban

Master's thesis presented for the double degree *Machine Learning and Data Mining* at Jean Monnet University and the IT department, major *Image, Modélisation et Informatique*, at CPE Lyon

in the *Methods and Computational Anatomy* team at Institut de Neurosciences de la Timone, at Aix-Marseille University

June 2021

Abstract

In our study, we are interested in cerebral connectivity mappings concerning U-shaped fibers located between the precentral gyrus and the postcentral gyrus, in the central human sulcus. The central sulcus is the anatomical border between the frontal lobe, approximately corresponding to the primary motor regions, and the parietal lobe, corresponding to the primary sensitive regions. The interest of these mappings is to understand the role of short U-shaped fibers, which correspond to 90% of the association fibers of the human brain, in this area in particular.

The establishment of these connectivity mappings was done in *L'étude de la connectivité structurelle des faisceaux d'association courts de la substance blanche du cerveau humain en IRM de diffusion* led by Alexandre Pron at the Intitut de Neurosciences de la Timone. It is based on diffusion-weighted magnetic resonance imaging (dMRI), followed by tractography and filtering of the tractograms which make it possible to select the candidate fibers. Only the fibers of the area of interest are kept and we proceed to a projection of it in a space where we apply a Gaussian kernel to obtain a connectivity mapping.

Interpreting connectivity mappings is difficult because there is significant variability between individuals. The challenge of our work is to improve the group representation and to allow a better equivalence between the latter and the individual subjects. To do this, we use the optimal transport theory. This theory is interesting in our case because the Wasserstein metric is more faithful between our subjects and therefore to build a more representative barycenter and to study the variety of our subjects. We also use the optimal transport matrix in order to match two subjects to each other, which is useful in the case of a label transfer from the group subject to an individual subject, and vice versa.

Keywords: Optimal transport, Wasserstein distance, Central sulcus, white matter, U-shaped fibers, connectivity mappings

Acknowledgements

I would like to thank Olivier Coulon for welcoming me to his team during this particular period, and for listening and giving sound advice. I would also like to thank the Centre de Calcul Intensif d'Aix-Marseille for access to its high performance computing resources which enabled us to carry out our experiments.

I am grateful to my academic tutors, Marion Foare and Marc Sebban, for their availability and listening during these last months, and also to the two teaching teams of my training and the three administrative teams of the UJM, CPE Lyon and the AMU without whom this internship could not have taken place. I sincerely hope that such a partnership will be repeated with other students.

I have a special thought for my little brother Rémy whom I made use of for the rereading of the following report, as well as to my close friends, with whom I shared the adventures of this internship, in particular with Clémence for encouraging and supporting me during the past months.

Contents

Introduction **1**

1 Theoretical framework and presentation of the study **3**
1.1 White matter fibers . . . 3
1.1.1 Introduction . . . 3
1.1.2 Discrete connectivity mapping . . . 4
1.1.3 Continuous connectivity mapping . . . 5
1.2 Optimal transport . . . 7
1.2.1 Formulation of the problem . . . 7
1.2.2 Wasserstein distance . . . 8
1.2.3 Regularisation . . . 9

2 Methodology **10**
2.1 Establishment of group subjects . . . 10
2.1.1 2-Wasserstein distance . . . 10
2.1.2 Group subject . . . 11
2.1.3 Information on group subjects . . . 13
2.2 Structure within the subject set . . . 13
2.2.1 Clustering by K-medoids . . . 13
2.2.2 Dimension reduction by isomap . . . 14
2.3 Backpropagation of the group subject to an individual subject . . . 15
2.3.1 Simulation of discrete subjects . . . 16
2.3.2 Visualisation tools . . . 16
2.3.3 Optimal transport matrix . . . 16

3 Results and analysis **18**
3.1 Establishment of group subjects . . . 18
3.1.1 Experiment 1: Mean . . . 18
3.1.2 Experiment 2: Centroid . . . 18
3.1.3 Experiment 3: Barycenter . . . 19
3.1.4 Conclusion . . . 21
3.2 Structure within the subject set . . . 22
3.2.1 Subject clustering . . . 22
3.2.2 Isomap . . . 25
3.2.3 Experiment 1: Isomap in dimension 2 . . . 26
3.2.4 Experiment 2: Isomap in dimension 1 . . . 27
3.2.5 Experiment 3: Sliding Barycenter . . . 28
3.2.6 Conclusion . . . 30
3.3 Correspondence between group subjects and individual subjects . . . 30
3.3.1 Gaussian mix to Gaussian cloud . . . 30
3.3.2 Entropy regularisation - discrete mappings . . . 31

3.3.3 Entropy regularisation - continuous mappings 33
3.3.4 Simulation of real data 33
3.3.5 Conclusion 35

Conclusion **36**

Bibliography **37**

Appendices **37**
3.4 Centroid 37
3.5 Isomap 39

List of Figures

1.1 homunculus . . . 3
1.2 Samples of discrete connectivity mapping . . . 5
1.3 Samples of Continuous connectivity mapping . . . 5
1.4 Clustering . . . 6
1.5 Backpropagation . . . 6
1.6 Monges problem of déblais and remblais . . . 7

3.1 Group representation: summation of subjects - both hemispheres . . . 18
3.2 Centroids - both hemispheres . . . 19
3.3 W_2^2 as a function of the number of points - both hemispheres . . . 19
3.4 Group representation: barycenter of subjects . . . 20
3.5 Silhouette and elbow score - L . . . 22
3.6 Silhouette and elbow score - R . . . 22
3.7 Reorganised W_2^2 matrices - L . . . 23
3.8 Reorganised W_2^2 matrices - R . . . 23
3.9 Sub-cluster barycenter k = 2 - both hemispheres . . . 24
3.10 Sub-cluster barycenter k = 3 - both hemispheres . . . 25
3.11 Isomap dim 2 - L . . . 26
3.12 Isomap dim 2 - R . . . 26
3.13 Isomap dim 1 - L . . . 27
3.14 Isomap dim 1 - R . . . 27
3.15 Hand knob . . . 28
3.16 Samples of sliding barycenter - L . . . 29
3.17 Samples of sliding barycenter - R . . . 29
3.18 Position of the maximums of the central area . . . 30
3.19 Gaussian mixture to Gaussian cloud . . . 31
3.20 Gaussian mixture to Gaussian cloud - continuous case . . . 31
3.21 OT matrix - entropic regularisation . . . 32
3.22 Source projected - entropic regularisation - discrete case . . . 32
3.23 Source projected - entropic regularisation (reg = 100) - continuous case . . . 33
3.24 Entropic regularisation (reg = 10) - continuous case . . . 34
3.25 Entropic and Tikhonov regularisation (reg1 = 10, reg2 = 10) - continuous case . . . 35
3.26 Isomap dim 2 - L . . . 39
3.27 Isomap dim 2 - R . . . 40
3.28 Isomap dim 1 - L . . . 41
3.29 Isomap dim 1 - R . . . 42

List of Tables

1.1 Information on data for the number of points 4

3.1 Position of local maximum - summation group representation 18
3.2 Position of local maximum - barycenter group representation 20
3.3 Comparison of the barycenters with the 100 subjects and each other - R 20
3.4 Comparison of the barycenters according to the order of the subjects - R 21
3.5 Position of local maximum - 2 clusters 23
3.6 Position of local maximum - 3 clusters - both hemispheres 24
3.7 Gaussian mixture to Gaussian cloud - discrete case 30
3.8 Simulation of real data 34
3.9 Centroid - distance of one subject from the other 100 - left hemisphere 37
3.10 Centroid - distance of one subject from the other 100 - right hemisphere 38

Introduction

The object of this thesis is the study of the connectivity of the U-shaped white matter fibers, in the central sulcus, from discrete density maps of each hemisphere of the 100 available subjects. The central sulcus divides the parietal lobe, it separates in particular the precentral gyrus corresponding to primary motor skills and the postcentral gyrus corresponding to the somato-sensorial areas. It is therefore interesting to know the organisation of short fibers in this region in order to better understand how those areas communicate.

MeCa

The Institute of Neuroscience of Timone attached to Aix-Marseille University aims to study neurosciences in depth in order to better understand the central nervous system. Different teams are organised around this theme, whether in the field of cognitive neurosciences, neurophysiology, cellular or behavioral neurosciences. Grouped in the same building located in the heart of the Faculty of Medicine, near CHU Timone, common means are made available to common resources in order to carry out the various experiments. We have in particular the presence of an animal facility, a mechanical and electronic workshop or even imaging tools available.

It is in this context that the "Méthodes et Anatomie Computationnelle" (MeCA) team fits. The goal of the team is to study the variability of the cortex and the connectivity of white matter in the brain from MRI acquisition on human and non-human primates. Various tools have been developed within this team and made available in particular within the BrainVisa project which is a neuroimaging software.

Internship tasks

The main goal of this internship is to improve the results of the segmentation made in *l'étude de la connectivité structurelle des faisceaux d'association courts de la substance blanche du cerveau humain en IRM de diffusion*. This segmentation made on a group representation is back-propagated to the separate subjects before being put back into a parameterised space representing the connections between two brain areas. In our case, we are interested in the earlier results used before the segmentation process on the group representation. We work on two levels: the construction of a better group representation, and a better correspondence between the group representation and the separate subjects.

In the case of the construction of a group representation, we used the metric of 2-Wasserstein in order to do our manipulations on the space of our data. This metric allows us to define a geodesic distance between subjects. Using this distance we study different group representation propositions and also the structure of our data in the sense that we are looking for population subgroups or failing that an axis of variation.

Moreover, in the case of the mapping between the group representation and the associated subjects the optimal transport is used. This method of having the most optimal result to transform one distribution into another, in our case a group connectivity mapping to another individual connectivity mapping.

Thesis organisation

This report is organised into 4 chapters: the theoretical framework and presentation of the study, which makes it possible to present the context in which our thesis takes place; the methodology, where we set the necessary prerequisites for our various experiments; the results and analysis, where we present the different experiments that we have done and grouped according to the tasks mentioned in the previous section and a general conclusion, which allows us to summarise what we have been able to accomplish and a discussion of areas for improvement.

The different works are grouped into three categories: the creation of the group representation, the study of an organisation within the population and finally the correspondence between the group representation and the associated separate subjects. It is important to note that some of this work is still in progress when this report is submitted.

Chapter 1

Theoretical framework and presentation of the study

1.1 White matter fibers

1.1.1 Introduction

In the case of the study of the white matter fibers connectivity, we are interested in the central sulcus. The central sulcus marks the separation between the frontal lobe and the parietal lobe. On both sides of the sulcus, we have the primary motor cortex, at the frontal lobe; and the primary somatosensory cortex, also known as the primary sensory cortex, at the parietal lobe. We study the connectivity in that zone in order to point out the interactions between both areas. We can associate a homunculus, which is a deformed representation of the human body, for each of the two areas (figure 1.1).

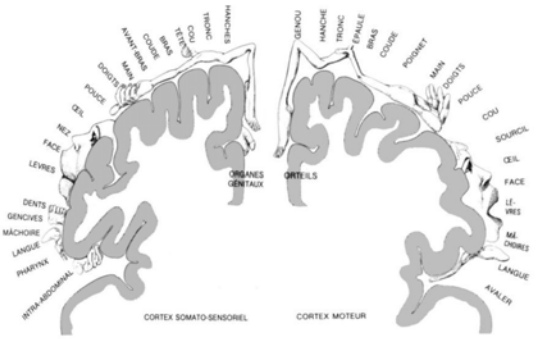

Figure 1.1 – homunculus

We observe a great proximity between the two homunculus of the primary sensory cortex and the primary motor cortex. We note that the feets are located in the upper zone and the mouth in the lower zone, whereas the hands are in the central position.

The connection between those two areas is made by the axons of the neurons. An axon is the extension of a neuron towards another neuron. Also named as fiber, it allows the transmission of information. In the case of our study, we are particularly interested in U-shaped short fibers.

Mentioned for the first time in 1885 by Meynert after dissections, it is only until 2012 that we get the first mapping of the short association fibers of the frontal and parietal lobes by Catani. It is important to note that such mappings are based on diffusion MRIs (DW-MRI).

The diffusion MRIs method relies on the indirect determination of fibers. As such, without dissections, we are unable to tell with certitude that the fibers represented by the mapping represent an anatomical reality.

There are a great number of descriptions of the U-shaped short fibers. We retain the definition proposed by Pron which relies on the one of Meynert and Dejerine. We consider short fibers of a length inferior to 30mm, located between two consecutive gyri. A gyrus being a ridge, it corresponds to the ridges on both sides of the central sulcus that we study.

1.1.2 Discrete connectivity mapping

It is in that context that our study is part of. We recover the connectivity mapping of 100 subjects for each of the two hemispheres, generated from the work done by Alexandre Pron in the *Etude la connectivité structurelle des faisceaux d'association courts de la substance blanche du cerveau humain en IRM de diffusion*[Pro19]. The generation of those connectivity mapping is described below.

With the results of the diffusion MRI, we proceed to a tractography which allows us to determine the potential fiber beams, also known as streamlines. This method relies on diffusion tensors. Afterwards, we apply a filter on the streamlines based on different criterias in order to reduce the number of candidate fibers and to get as close as possible to the anatomical reality. We use the previously provided definition of a short fiber, spatially restraint around the central sulcus. Then we do a parameterisation of the space depending on the ridge lines of the gyri from 0 to 100 between two ends manually défined for each subject. This parameterisation is based on anatomical landmarks, which are low-variancy points from one to another. We use another central landmark in order to better align the mapping. We then have our connectivity mapping of the two hemispheres for 100 subjects.

It is important to note that after the processus we have discrete connectivity mapping (figure 1.2). Each connectivity mapping contains points of $\mathbb{R}^2$. For each axis, x corresponds to the precentral lobe and y corresponds to the postcentral lobe, The value 0 is assigned to the ventral level while the value 100 is assigned to the dorsal level. We have the following characteristics for our 100-subject dataset in the table 1.1.

hemisphere	left	right
min	943	582
median	2409	2008
max	4468	4029
mean	2498	2087
standard deviation	700	658
Excess Kurtosis	-0.23	0.17

Table 1.1 – Information on data for the number of points

Kurtosis is the fourth statistical moment, defined as :

$$Kurt(X) = \frac{E([X-\overline{X}]^4}{E([X-\overline{X}]^2)^2}$$

We use Fisher's definition in our case, i.e. we subtract 3 at this point. The closer it is to 0, the more the distribution of values tends towards a Gaussian distribution. This is the case with our data.

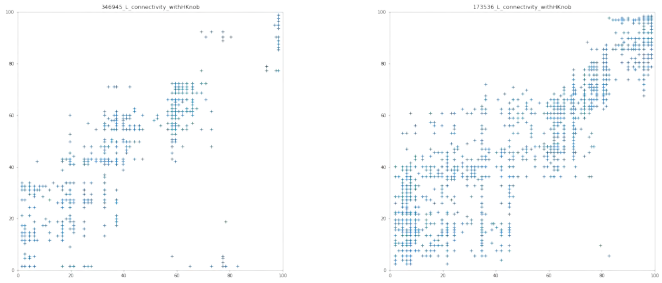

Figure 1.2 – Samples of discrete connectivity mapping

Of the 100 study subjects we have, there is a large variability in the connectivity mapping between them for a given hemisphere in terms of number of points and distributions.

1.1.3 Continuous connectivity mapping

In order to better understand these differences in distributions, we prefer to do the analysis on continuous connectivity mappings. We convolve by a Gaussian kernel to obtain a continuous mapping from our discrete mappings.

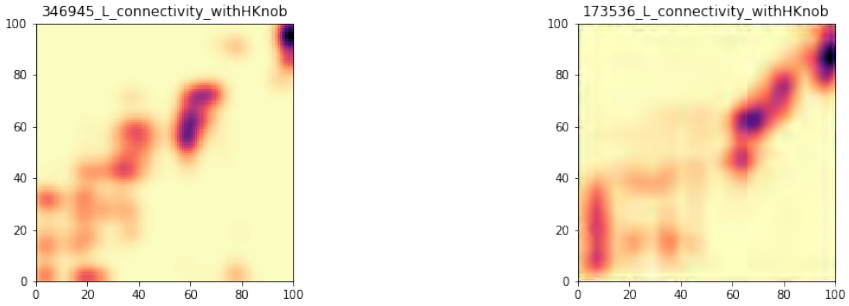

Figure 1.3 – Samples of Continuous connectivity mapping

Through the two example subjects of figure 1.3, we notice that the marginal law according to the x-axis corresponds to the distribution of fibers in the precentral zone, motor area, whereas that according to the y-axis corresponds to the post-central zone, sensory area. The distributions are mainly concentrated around the main diagonal. This is because the closer one is to this diagonal, the smaller the distance between the motor area and the sensory area. Furthermore, we can see that there are three major sets: two at the ends of the diagonal and one in the centre. By taking again the representation of the homunculus in figure 1.1, we have the zone in bottom on the left which corresponds to the mouth, that in the centre to the hand and finally that in top on the right to foot. These densities can be approximated by a Gaussian juxtaposition.

One of the main issues at this stage is the establishment of the group topic. In the case of Pron[Pro19]'s thesis, he studied the group representation resulting from the summation of the distinct subjects, in other words we have the average of the subjects to within one normalisation. Once this group representation is generated, we try to isolate features such as the fingers of the hand on it. This corresponds to Gaussian groupings. To do this, we perform a segmentation on the continuous connectivity mapping of the group representation. We use the algorithm Density-based spatial clustering of applications with noise (DBSCAN) (figure 1.4).

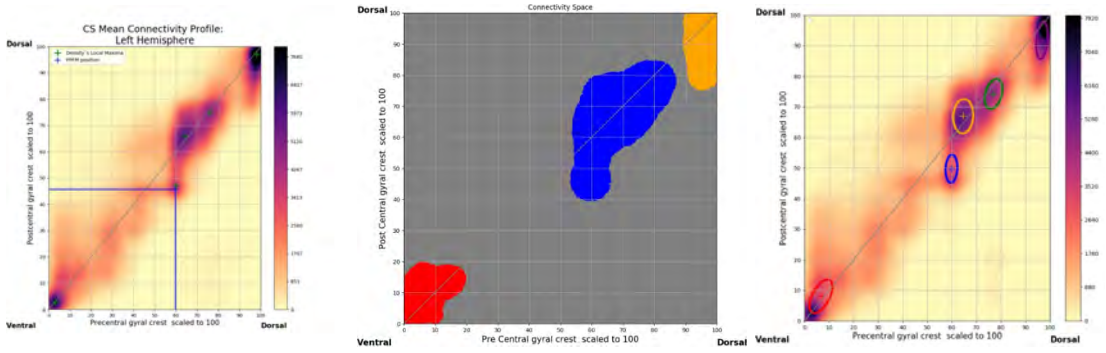

Figure 1.4 – Clustering

We backpropagate the segmentation from the continuous mapping to the discrete mapping and then from the discrete to the parameter space to match the candidate U-fibers (figure 1.5).

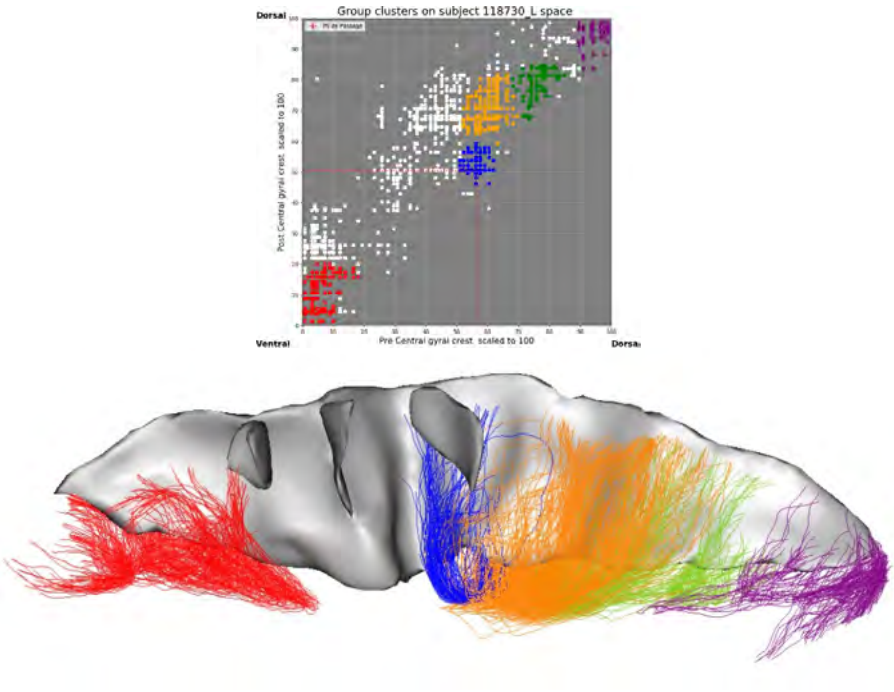

Figure 1.5 – Backpropagation

The goal of our work is to improve the segmentation result in the parameterised space. For that we have three axes of work:

— improve the quality of the group representation so that it is more representative of the dataset;
— determine if there is any organisation within our data so that we can define subgroups or determine an axis of variability from one subject to another;
— have a better back-propagation between the segmentation on the continuous mapping to the discrete mapping.

In order to solve these problems, we are particularly interested in optimal transport. This is detailed in the next section.

1.2 Optimal transport

Optimal transport is the transformation of one distribution into another with respect to a given cost function. In other words, it is a method to define the best possible transformation to match two distributions. There are two main formulations of this problem and like any optimisation problem it has a constrained formulation and different methods of solution.

1.2.1 Formulation of the problem

We present two formulations of the optimisation problem to determine the optimal transport solution.

Monge's formulation

The first formal formulation of the optimal transport problem was published in *Mémoire sur la théorie des déblais et des remblais* in 1781 by Gaspard Monge[Mon81]. The problem consists of the transport of a certain quantity of earth from the point of extraction to the point of use.

The aim is to minimise the energy required to carry out the operation, i.e. both the transport and the quantity previously extracted from the mine and the assignment to the site of use, by minimising the points of storage and movement between the two sites. The problem can be illustrated[Céd09] in figure 1.6.

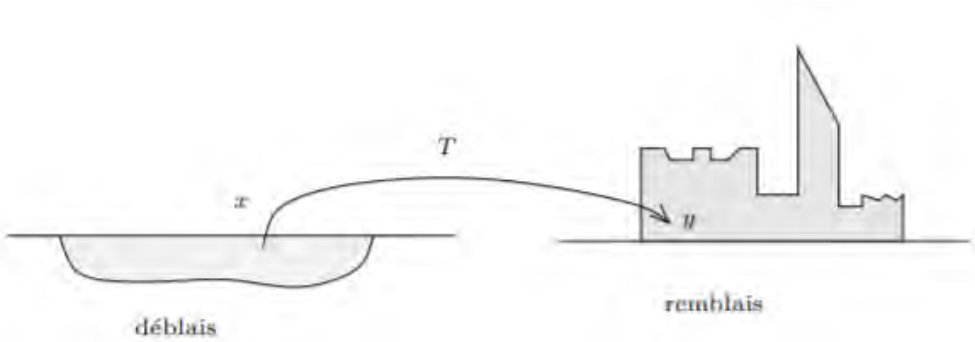

Figure 1.6 – Monges problem of déblais and remblais

We note T the optimal transport application allowing to transfer a certain volume of excavation from position x, noted $\mu_s(x)$ with s for source, towards a certain necessary volume of embankment at position y, noted $\mu_t(y)$ with t for target. In a more formal way we have:

$$\mu_t(y) = \mu_s(x) = \mu_s(T^{-1}(y))$$

The transfer cost between $x \in X$ and $y \in Y$ is defined as:

$$c(x, y) = c(x, T(x))$$

$$w(x, y) = \mu_s(x)c(x, y) = \mu_s c(x, T(x)).$$

The elementary energy for the displacement of one unit of elementary volume is deduced:

$$W = \int_X \mu_s(x)c(x, T(x))dx = \int_X c(x, T(x))\mu_s(dx)$$

We try to minimise this necessary energy according to the transport, we note MP the set of transports of the associated space preserving the measure of μ_s and μ_t also noted $T * \mu_s = \mu_t$. We have the following minimisation problem which is the Monge formulation:

$$W = min_{\tilde{T} \in MP} \int_X c(x, \tilde{T}(x)) \mu_s(dx) \tag{1.1}$$

It is important to note that in Monge's formulation we assume that the μ_s and μ_t distributions are continuous. Moreover, we can bring back all the previous definitions in the case of the formalism of probabilities, subject to a normalisation of the distributions to have densities of probability.

We have $T = argmin_{\tilde{T} \in MP} \int_X c(x, \tilde{T}(x)) \mu_s(dx)$, the optimal transport application which allows to minimise the cost function.

Kantorovich's formulation

Kantorovich's work is described as *The birth of the modern formulation of optimal transport*[Céd09]. In this section we focus on the reformulation of the Monge problem:

$$\begin{aligned} W(\mu_s, \mu_t) &= min_{\pi \in \Pi(\mu_s, \mu_t)} \int_{X \times Y} c(x, y) \pi(x, y) d(x, y) \\ &= min_{\pi \in \Pi(\mu_s, \mu_t)} \int_{X \times Y} c(x, y) d\pi(x, y) \\ &= min_{\pi \in \Pi(\mu_s, \mu_t)} < M_\pi, M_c > \end{aligned} \tag{1.2}$$

By adapting the usual formalism[Cav19] with our notations we have $\Pi(\mu_s, \mu_t)$, the coupling designating the set of probabilities π on $X \times Y$ having marginal laws μ_s and μ_t. We notice that we find this idea of energy minimisation in Kantorovich's formula by posing M_π the matrix associated to the π application and M_c, the matrix associated to the c application.

We choose not to go into detail, especially for the questions of existence and uniqueness for the resolution of the optimal transport. Moreover, this formulation has the advantage of being able to solve the transport problem in a more general framework without any continuity condition on the associated probability densities.

1.2.2 Wasserstein distance

Throughout the different formulations for establishing the optimal transport solution, we have not yet characterised the cost function c, which quantifies the measurement gap between the source and target distribution. Given c as a usual distance in space, we define the Wasserstein distance as:

$$W_p^p(\mu_s, \mu_t) = min_{\pi \in \Pi(\mu_s, \mu_t)} \int_{X \times Y} d_p(x, y)^p d\pi(x, y) \tag{1.3}$$

We thus define the Wasserstein distance, with $p >= 1$ and d the usual norm of the space $\mathbb{R}^p$, while the p-Wasserstein distance is defined as follows:

$$W_p(\mu_s, \mu_t) = min_{\pi \in \Pi(\mu_s, \mu_t)} [\int_{X \times Y} d_p(x, y)^p d\pi(x, y)]^{1/p} \tag{1.4}$$

The minimum Wasserstein distance is reached when the optimal application, noted T, between the source and the target is determined. The overall expenditure can be found by the following formula:

$$W_p(\mu_s, \mu_t) =< T_{\mu_s, \mu_t}, D_p^p(\mu_s, \mu_t) >^{1/p} \tag{1.5}$$

For example, for $p = 2$ with $(x \in X) \subset \mathbb{R}^2$, $(y \in Y) \subset \mathbb{R}^2$ and d the Euclidean norm of $\mathbb{R}^2$, we have:

$$\begin{aligned} W_2^2(\mu_s, \mu_t) &= min_{\pi \in \Pi(\mu_s, \mu_t)} [\int_{X \times Y} \|x, y\|_2^2 d\pi(x, y)] \\ &= min_{\pi \in \Pi(\mu_s, \mu_t)} < \pi_{\mu_s, \mu_t}, D_2^2(\mu_s, \mu_t) > \end{aligned}$$

1.2.3 Regularisation

The resolution of minimisation problems can be constrained in order to facilitate convergence and/or reduce the complexity of the resulting model[Rot20], expressed as follows:

$$W_p^p(\mu_s, \mu_t) = min_{\pi \in \Pi(\mu_s, \mu_t)} < \pi_{\mu_s, \mu_t}, D_p^p(\mu_s, \mu_t) > + \epsilon f(\pi(\mu_s, \mu_t)) \tag{1.6}$$

In our case, we are particularly interested in the regularisation of max entropy[PC20]. We can define f as the Kullback-Leibler divergence:

$$f(\pi(\mu_s, \mu_t)) = KL(\pi(\mu_s, \mu_t) | \pi(\mu_s, \mu_t)^\epsilon) = \int_{X \times Y} \pi(\mu_s, \mu_t) log(\frac{\pi(\mu_s, \mu_t)}{\pi(\mu_s, \mu_t)_\epsilon})$$
$$\text{with } \pi(\mu_s, \mu_t)_\epsilon = e^{-\frac{1}{\epsilon} D_p^p(\mu_s, \mu_t)}$$

We can also use the Tikhonov regularisation which allows us to put a condition on the amplitudes of the function which we regularise[Pro05]. We define f as the associated norm of the p-dimensional space, where p is the same dimension as the Wasserstein distance. In the case of matrix calculus, we use the frobenius norm. We pose:

$$f(\pi(\mu_s, \mu_t)) = \|\pi(\mu_s, \mu_t)\|_p^p$$

In our case we will use $p = 2$. We can cumulate these different regularisations between them if necessary.

Chapter 2

Methodology

2.1 Establishment of group subjects

The construction of a group subject allows for a subject that is representative of all subjects. The aim is to establish the least biased group subject possible in order to extract common characteristics within the subject population. There are various procedures for obtaining this, such as

- the average of the subjects that was determined in the previous work [Pro19];
- the centroid subject, in the sense of the 2-Wasserstein distance, which is the most representative subject of the set of subjects;
- the barycenter of the subjects, in the sense of the 2-Wasserstein distance, which is based on the Fréchet mean with the 2-Wasserstein distance in our case.

For the interpretations and visualisation of the different proposals, the continuous connectivity mappings are presented, but the discrete connectivity mappings are used to generate these group topics.

2.1.1 2-Wasserstein distance

The theory of optimal transport was presented in chapter 1, in this chapter we explain the specific averages used to construct a group subject. In the case of the optimal transport formalism, we are only interested in the 2-Wasserstein distance, noted W_2, and not in the optimal transport matrix noted T.

Cost matrix

The cost matrix is defined as the matrix of all the distances of the points between the source and target subjects. In our case the cost matrix is determined according to the Euclidean distance to the square in order to have the Wasserstein distance with $p = 2$. The Euclidean distance seems to us to be appropriate in our case insofar as the coordinates of the points are in $\mathbb{R}^2$ and that the usual norm for this space is the Euclidean norm. This cost matrix is calculated by the function spatial.distance.cdist, from the library scipy[Vir+20].

Wasserstein distance

The 2-Wasserstein distance is the metric of our data space. Let T be the optimal transport matrix, M the cost matrix, we define the 2-Wasserstein distance between two subjects as follows:

$$W_2 = \sqrt{< OT, M >}$$

Between N subjects, like the cdist function in the scipy library but with the Wasserstein distance, we use the following algorithm:

Algorithm 1 cdist_wasserstein

Input:
L_coordinate ▷ List of points'coordinates for N subject
L_weight ▷ List of associated weights for N subject
Output: cdist ▷ Output: Matrix of wasserstein's distances
procedure CDIST_WASSERSTEIN(L_coordinate, L_weight):
  $Mat_{distance} \leftarrow 0_{N,N}$ ▷ initialisation matrix of distance
  for $i \leftarrow 1$ *to* N **do**
    $x_s \leftarrow L\_coordinate[i]$ ▷ point's coordinates of current source
    $w_s \leftarrow L\_weights[i]$ ▷ the weights associated
    for $j \leftarrow 1$ *to* N **do**
      $x_t \leftarrow L\_coordinate[j]$ ▷ point's coordinates of current target
      $w_t \leftarrow L\_weights[j]$ ▷ the weights associated
      $M \leftarrow distance(x_s, x_t)$ ▷ compute cost matrix
      $OT \leftarrow function(w_s, w_t, M)$ ▷ compute OT matrix
      $Mat_{distance}[i,j] \leftarrow \sqrt{< OT, M >}$ ▷ distance
  $cdist \leftarrow Mat_{distance}$

Optimal transport matrix

The optimal transport matrix is defined as the matrix minimising the Wasserstein distance for a given cost matrix. In the case of this chapter, this matrix is determined by the Earth Mover's distance algorithm implemented in the POT library. This algorithm needs as input the set of weights associated with the point of each distribution, source and target, as well as the associated cost matrix. There is also an output condition depending on the number of iterations.

2.1.2 Group subject

Average

In the previous work, in order to define the group subject, the sum of all subjects for each of the two hemispheres is calculated[Pro19]. This corresponds to the average of all the data with one normalisation. This group connectivity mapping from the average of the subjects for each hemisphere is sought.

Centroid

One way to determine the group subject is to choose the subject that is most representative of the set of subjects. That is to say, the one that minimises the differences between each subject, in other words the most central subject in our subject space. We use the function cdist_wasserstein which allows us to calculate the set of distances on our 100 subjects for a given hemisphere. Once the matrix of distances between all subjects is established, we calculate for each given subject the sum of its distances to the square and we recover the subject corresponding to the smallest of the distances. The centroid is then obtained.

With the result from the function cdist_wasserstein, we can rank the different subjects

in order from the most representative to the least representative. To do this, the 2-Wasserstein distance is used to calculate the average distance required for a given subject to move from it to the other 100 subjects. The set of distances to the square is summed up two by two. The average distance is defined as:

$$\overline{W_2} = \sqrt{\frac{1}{N}\Sigma W_2^2}$$

In order to better perceive the differences, one will use in a joint way this distance to the square, without the normalisation which one notes ΣW_2^2.

We determine the centroid subject and determine whether there is a relationship between the number of continuous points in the discrete connectivity mapping for each subject for a given hemisphere and the 2-Wasserstein distance.

Barycenter

There is a generalisation of the notion of average for non-Euclidean spaces to respect the geodesy of our problem. This generalisation is called the Fréchet mean. It is based on a definition of the most suitable distance for the space under study. In our case we use the 2-Wasserstein distance to the square to establish this average.

$$\Psi(\mu) = \sum_{i=1}^{N} W_p^p(\mu, \nu_i) \text{ [CD14]}$$

The Wasserstein barycenter is defined as follows:

$$m = argmin_{\mu \in (\mathbb{R}^2)^k}(\sum_{i=1}^{N} W_p^p(\mu, \nu_i))$$

We use the following algorithm[WRT18a] to determine this solution:

Algorithm 2 $2 - Wasserstein\_Barycenter$

Input:
L_coordinate ▷ List of points'coordinates for N subjects
L_weight ▷ List of associated weights for N subjects
L_contribution ▷ List the contribution of each to the barycenter
X and a ▷ Support of barycenter
Output: X (size k,d) ▷ Barycenter
procedure $2 - Wasserstein\_Barycenter$(L_coordinate,L_weight,L_contribution,X,a):
  while X and a have not converged **do**
    $T_i \leftarrow 0_{k,d}$
    for $i \leftarrow 1$ *to* N **do** ▷ compute for each subject
      $M \leftarrow distance(X, L\_coordinate[i])$ ▷ compute cost matrix
      $OT \leftarrow function(a, L\_weight[i], M)$ ▷ compute OT matrix
      $T_i = T_i + L\_contribution[i].diag(a^{-1}).OT.M$ ▷ compute barycenter
    $X \leftarrow T_i$

Our aim is to build the barycenter on a topic-by-topic basis. This allows us to update the barycenter in case we have a new subject in the future. Moreover, opting for an iterative construction has the advantage of saving computational resources as we can start from the previous step. We have therefore adapted the free_support_barycenter algorithm in the POT library.

We try to show the impact of the initialisation support for the calculation of the barycenter, as well as the order of presentation of the subjects. This analysis allows us to check the robustness of our barycenter construction. We take a precision of 10^{-4} for the calculation of the optimal transport matrix. Given the way discrete connectivity mappings are generated, we need a distance precision of 10^{-1}, so we will display the results at 10^{-2}.

2.1.3 Information on group subjects

In order to compare our different group subjects, we use ΣW_2^2 and $\overline{W_2}$. The smaller these two measures are, the more representative the subject is of all subjects. We also use the position of the remarkable local maximums to characterise our group subjects. We provide the coordinates and the associated continuous connectivity mapping.

2.2 Structure within the subject set

Our aim is to determine whether there are subgroups of populations within our subjects, and if so to characterise them. To do this, we use the K-medoids algorithm which allows us to determine clusters within our dataset, while a dimension reduction algorithm allows us to have a lower dimensional representation of our dataset and to judge the relevance of the defined clusters. Care is taken to use or adapt algorithms so that they can use the 2-Wasserstein distance or the distance matrix of the 100 subjects previously determined.

2.2.1 Clustering by K-medoids

We use the K-medoids algorithm which is a derivative of the K-means algorithm also known as the partition algorithm around medoids (PAM)[KR87]. This algorithm makes it possible to use metrics other than Euclidean ones, which is interesting in our case with the 2-Wasserstein distance. Contrary to K-means which calculates the barycenter to qualify the cluster, here we look for the centroid of the cluster, which hence is an element of our set. This method ensures that the central subject determining the cluster belongs to the space of elements. In other words, we better respect the geodesy of the problem by choosing the centroid rather than the barycenter in our case. The result is not unique because it depends on the initialisation of the K-centroid.

In the case where we provide directly the distance matrix between the subjects, it has the advantage of not needing to recalculate the distances between the different subjects to determine the centroid of the cluster because these are already present in the distance matrix. There are several scores to determine the quality of a clustering. In our case we use the elbow score and the silhouette score. These scores make it possible to determine the best number of clusters and also the quality of those. Moreover, the K-medoids algorithm provides a result depending on the initialisation points for each cluster. In order to minimise the impact of this initialisation, the algorithm is re-run several times.

Elbow score

This score is used to determine the best number of clusters if we determine an inflection point of the curve. This score represents the sum of the distances between an element of a cluster and its defining center[Sat+04]. This score gives an idea of the compactness of a cluster.

Silhouette score

This score[Rou87] translates the separability of the clusters, i.e. to what extent they are distinct. The closer the score is to 1, the more distinct the clusters are. This score is determined in the following way: $s = \frac{b-a}{max(a,b)}$ with a, the average intra-group distance and b the average distance closest to the group for each sample.

The set of these scores allows us to determine the optimal number of clusters in the case that it exists. We have no knowledge of the expected number of clusters nor of the existence of subgroups. Moreover, it allows us to determine the best solution among the n repetitions in order to minimise the impact of the initialisation topics. We have the following algorithm:

Algorithm 3 n_Kmedoids

```
Input:
cdist                                          ▷ matrix of distance
nb_cluster                                     ▷ number of cluster
iter_max                                       ▷ number of global iteration
Output: labels                                 ▷ labels for each subject
procedure N_KMEDOIDS(cdist, nb_cluster, iter_max):
    ϵ ← +∞
    for i ← 1 to iter_max do
        l ← Kmedoids(cdist, nb_cluster)
        if elbow_score(cdist,labels)<ϵ then
            ϵ ← elbow_score(cdist, labels)
            temp ← l
labels ← temp
```

We seek to determine the ideal number of subgroups within our population and to verify the relevance of these. To do this, we reorganise the 2-Wasserstein distance matrix, and determine the barycenter of the subgroups and the associated local maxima for different values of k.

2.2.2 Dimension reduction by isomap

The isomap algorithm[TSL00] allows a representation of the data in lower dimension. This algorithm is based on Principal Component Analysis (PCA) and MultiDimensional Scaling (MDS). These two combined approaches allow to keep the geodesy of the problem at best in the way that we will reduce the dimensionality of the problem, role of the PCA, while preserving as much as possible the relations which the subjects have between them, role of the MDS.

Algorithm 4 Isomap

First step: Construct neighborhood graph
Define the graph G over all data points by connecting points i and j if i is one of the KNN of j. Set edge lengths equal to $d_x(i, j)$.

Second step: Compute shortest paths
Initialise $d_G(i, j) = d_x(i, j)$ if i, j are linked by an edge $d_G(i, j) = \infty$ otherwise. Then for each value of $k \in [\![1; N]\!]$ In turn, replace all entries $d_G(i, j)$ by $min(d_G(i, j), d_G(i, k) + d_G(k, j))$. The matrix of final values $D_G = d_G(i, j)$ will contain the shortest path distances between all pairs of points in G.

Third step: Construct dimensional embedding
Let λ_p be the path eigenvalue (in decreasing order) of the matrix $\tau(D_G)$, and v_p^i be the i-th component of the p-th eigenvector. Then set the p-th component of the d-dimensional coordinate vector y_i equal to $\sqrt{\lambda_p v_p^i}$

The hyperparameter corresponding to the number of neighbours is considered in several studies notably in *Selection of the optimal parameter value for the Isomap algorithm*[SMR06], and in *Selection of the Suitable Parameter Value for ISOMAP*[SH12].

We try to determine if there is a structure or organisation within our dataset, failing this, a variation axis using the isomap in dimension 1. To do this, we retrieve the labels obtained by the K-medoid and we associate them with the subjects in this new space. We also calculate the sliding barycenter in the space of the isomap of dimension 1 for all the subjects in order to highlight an axis of variation.

2.3 Backpropagation of the group subject to an individual subject

Our goal is to establish a process allowing us to transfer the segmentation from a group representation to the corresponding subjects in order to be able to study the specificities between the subjects. We can also work from labeled continuous or discrete connectivity mappings, at this point we have no preference given that DBSCAN labeled on a group mapping continues and that we backpropagate the labels to the associated discrete group mapping.

We are interested in the optimal transport in particular in the optimal transport matrix. This is equivalent to a contribution matrix between the points of a source towards the points of a target. In other words, it provides a correspondence between two subjects. In the case of this study we disregard the Wasserstein distance that we have used so far to focus on this optimal transport matrix.

In order to carry out this study, we are working on point cloud models in order to control the complexity of the problem. Visualisation tools are established to match points from a source subject to a target subject. We simulate the projection of labels associated with points from one subject to another. We do a study of the optimal transport matrix without and with regularisation.

2.3.1 Simulation of discrete subjects

We choose to model our real data by Gaussian mixtures. For this we use the POT library: Python Optimal Transport [Fla+21]to generate point clouds with a Gaussian distribution and we add as many clouds as there are subsets that we want in order to have our mixtures. Care is taken to have an equiprobable contribution from each point in the mixtures. The most advanced simulation we do is modeling the subsets identified by Pron. We reduce the spread of Gaussians, by changing the variance parameters, compared to the experimental results determined in order to succeed in identifying them more easily. We then introduce parasitic sets. We define the parasitic sets being sets not corresponding to any identification made by Alexandre Pron. For that we generate them in the same way as the standard sets, but with fewer points.

2.3.2 Visualisation tools

Matches

After solving the optimal transport problem between two subjects, we have an optimal transport and associated cost matrix. We choose to associate the label of a point of the source for each of the points of the target taking care that it is the source point with the greatest contribution which labels the corresponding target point. This allows us to have a surjective representation, that is to say that all the points of the target are labeled with at least one point of the source. To do this, one traverses the optimal transport matrix according to the axis corresponding to the points of the target, and we look for the point of the source whose point is the largest in this axis. A link can also be displayed between a target point and one or its source points in order to facilitate the visualisation of the projection. We can change the opacity of this link according to the contribution of the source point on the target.

Labels

To visualise the projection from the source to the target and simulate labels from the group representation, a label is associated with the different point of the source. This label set can be based on the position of the point relative to its neighborhood. For this, we generate a set of degraded color labels that we associate according to the position sorted in x and in y of the points of the subject.

We can also create a label set in relation to the contribution of each point within the subject, particularly useful in the case of a non-uniform distribution, which corresponds to an amplitude value in the case of a continuous connectivity mapping. To do this, we retrieve the distribution of the mapping, and we associate a degraded color label according to the distribution.

2.3.3 Optimal transport matrix

The essence of this study is based on the generation of its optimal transport matrices, without or with regularisation. There are several algorithms implemented in the POT library in particular the Sinkhorn algorithm which makes it possible to solve the problem of minimisation with an entropy regularisation.

If we explore the Tikhonov regularisation adapted to the 2-Wasserstein metric, we use the function ot.optim.cg, which does not admit entropy regularisation. It is also possible to

combine entropy regularisation and another regularisation with the function ot.optim.gcg.n axis of variation.

Chapter 3

Results and analysis

3.1 Establishment of group subjects

3.1.1 Experiment 1: Mean

Summing up the 100 discrete connectivity mappings for each hemisphere, we get the average group representation. It corresponds to a normalisation close to the average of the subjects. This generated group representation serves as a reference for the next experiments that we do (figure 3.1).

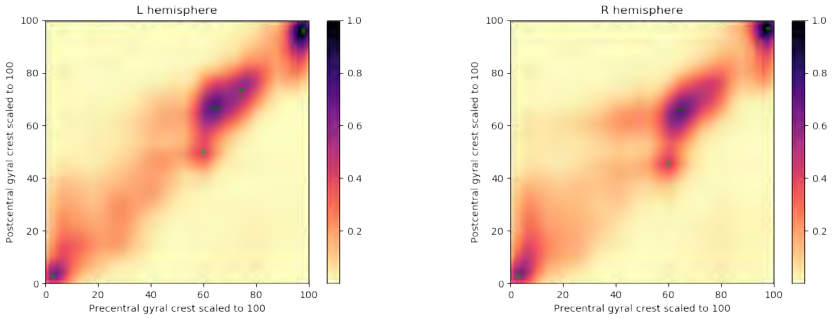

Figure 3.1 – Group representation: summation of subjects - both hemispheres

We determine the local maxima of interest by the method described by Alexandre Pron (table 3.1).

left	[96 98]	[74 74]	[67 64]	[50 60]	[3 3]
right	[97 98]		[66 65]	[46 60]	[3 3]

Table 3.1 – Position of local maximum - summation group representation

We see that the maximums allow us to characterise the clusters of interest which correspond lower left to the mouth, and upper right to the opposite foot to the concerned hemisphere side. This subject's 2-Wasserstein distance from the other 100 subjects was not calculated. because the number of points of the latter is too important so that it can be determined within reasonable calculation times.

3.1.2 Experiment 2: Centroid

The classification of the subjects according to the distances is in the appendix for the two hemispheres. The centroid subject for the left hemisphere is 495255 ($W_2^2 = 20508$) and for the

right hemisphere is 123117 ($W_2^2 = 22110$) (figure 3.2).

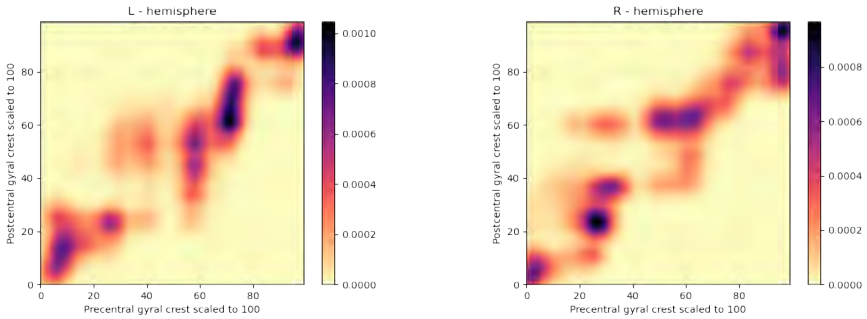

Figure 3.2 – Centroids - both hemispheres

We display the distance of 2-Wasserstein according to the number of points of the profiles (figure 3.3). We have details of the values in the appendix.

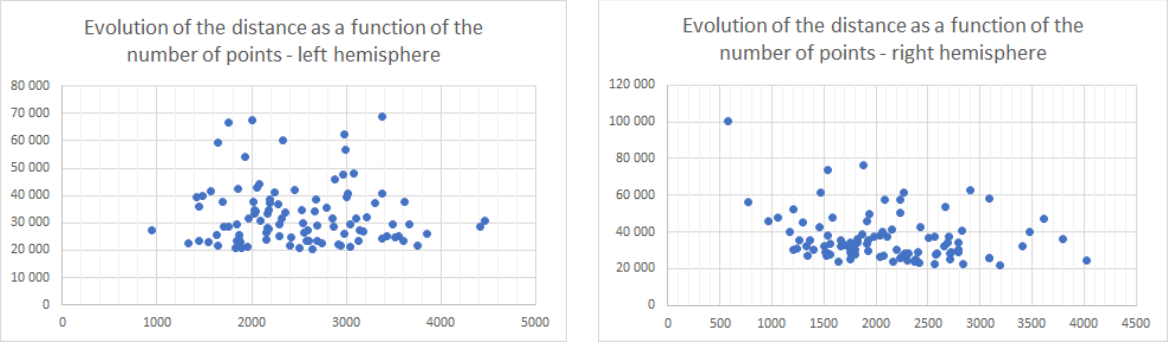

Figure 3.3 – W_2^2 as a function of the number of points - both hemispheres

There is no correlation between the number of points and the 2-Wasserstein distance for our dataset.

3.1.3 Experiment 3: Barycenter

Experiment 1: Determination of the barycenter

In this experiment we made the choice to calculate the barycenter in the order of the subjects defined during the previous experiment with the centroid. The most central subjects are added to the most extreme subjects. The centroid subject defined above is taken as starting support. We obtain the figure 3.4.

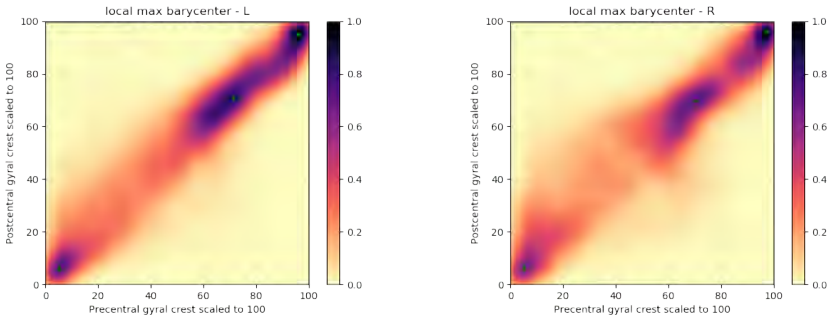

Figure 3.4 – Group representation: barycenter of subjects

We calculate the average 2-Wasserstein distance of this group representation needed to move from this one to another. We have the following value: $W_2^2 = 18494$ for the right hemisphere. We determine the position of the centers in the table 3.2:

left	[95 96]	[71 71]	[6 5]
right	[96 97]	[70 70]	[6 5]

Table 3.2 – Position of local maximum - barycenter group representation

We see that we do have a localised distribution around the main diagonal, which is expected. There is also an improvement in the representativeness of the subject compared to the centroid, because the average 2-Wasserstein distance is smaller. The calculation of the barycenter provides us with a better group representation.

Experiment 2: influence of the boot medium

We want to know if there is an incidence on the result of the barycenter according to the support we use for the initialisation. It is noted that the first subject conditions the number of points of the barycenter for the continuation. For this, we calculate the barycenter with different subjects drawn from our dataset. We take the subject containing the fewest points, the median, as well as the one containing the maximum. We also take a random subject, and a subject that we generated ourselves with the average number of points. We make the study arbitrarily on the right hemisphere, we consider that the conclusions are the same for the left hemisphere for the study of the barycenter. We get table 3.3.

Experiments		among the 100 subjects		among the experiences	
Name	Number of points	ΣW_2^2	$\overline{W_2}$	ΣW_2^2	$\overline{W_2}$
minimum	582	18576	13.7	12.51	1.17
random	768	18546	13.69	10.91	1.1
median	2040	18501	13.67	7.96	0.94
mean	2087	18500	13.67	7.91	0.93
centroid	3190	18494	13.67	7.63	0.92
maximum	4029	18492	13.67	7.4	0.9

Table 3.3 – Comparison of the barycenters with the 100 subjects and each other - R

Here we see a correlation between the number of points and the Wasserstein distance. The difference between the barycenters is very small or even negligible if we reduce these variations to the number of points of each. Furthermore, the incidence of the initialisation subject is

very low or even non-existent compared to the precision that we need. The differences in distances are explained more by the number of points than by the diversity of the barycenters generated. After passing the Gaussian kernel to generate the continuous mapping, we have very similar mappings. The number of points affects the resolution of the generated mapping but the position of the maximums remains unchanged.

Experiment 3: influence of the order of subjects

So far we have provided the subjects in the same order resulting from the calculation of the experiment to determine the centroid, we now study the impact on the order of presentation of the subjects in the establishment of the barycenter (table 3.4). We take the median subject as an initialisation support.

Experiment	among the 100 subjects		among the experiences	
	ΣW_2^2	$\overline{W_2}$	ΣW_2^2	$\overline{W_2}$
1	18499	13.66	8.91	0.99
2	18500	13.67	8.88	0.99
3	18504	13.67	9.50	1.02
4	18501	13.67	9.00	1.00
5	18498	13.66	8.91	0.99
6	18499	13.66	8.97	0.99
7	18499	13.66	8.80	0.98
8	18502	13.67	9.30	1.01
9	18501	13.67	9.15	1.00
10	18499	13.67	9.14	1.00

Table 3.4 – Comparison of the barycenters according to the order of the subjects - R

We observe that the order of presentation of the subjects has a greater impact than the support subject, but this influence is negligible if we reduce it to the number of points. By generating the continuous connectivity mapping these differences are small and can be controlled depending on the precision we need and the computation time available. Indeed, the determination of the coordinates of the local maximums remains unchanged.

3.1.4 Conclusion

The barycenter provides the best group subject in the direction that it minimises the distance of Wasserstein compared to other processes such as summation of subjects or the centroid. The barycenter is robust in the sense that it is invariant according to the support and the order of presentation of the subjects. The number of points required can be adapted according to the calculation time and the resolution of the continuous connectivity mapping. We choose to take the average number of points for all the subjects because it presents the best compromise in terms of calculation time and distance.

It is interesting to note that the intermediate result of calculating the centroid which is the distance of the subjects in relation to the others suggests that we can regroup the subjects in sub-groups. This hypothesis comes from the fact that we do not have the correlation between the number of points and the Wasserstein distance which implies that we have different underlying structures.

3.2 Structure within the subject set

3.2.1 Subject clustering

Experiment 1: Number of candidate subgroups

Our n _Kmedoids algorithm uses the scikit-learn-extra library which provides us with the K-medoids algorithm. We provide the 2-Wasserstein distance matrix as input and we determine the labels for $k = 2$ up to $k = 10$ which corresponds to 10 subjects on average per cluster. We choose to use the *k-medoids++* initialisation which corresponds to the initialisation described in the initial algorithm of PAM [Mar63]. Within this process we calculate the elbow score which represents the dispersion of values within the cluster, and the silhouette score which represents the separability of the clusters. We do 5000 iterations in order to determine the best label because for k=2 this corresponds to the scanning of the set of possible initialisations. We have 4950 possibilities for 2 initialisations among 100 while for k=3 we have 161 700 possibilities.

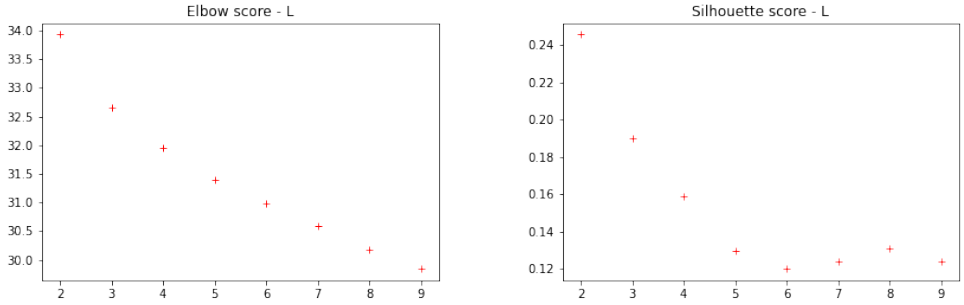

Figure 3.5 – Silhouette and elbow score - L

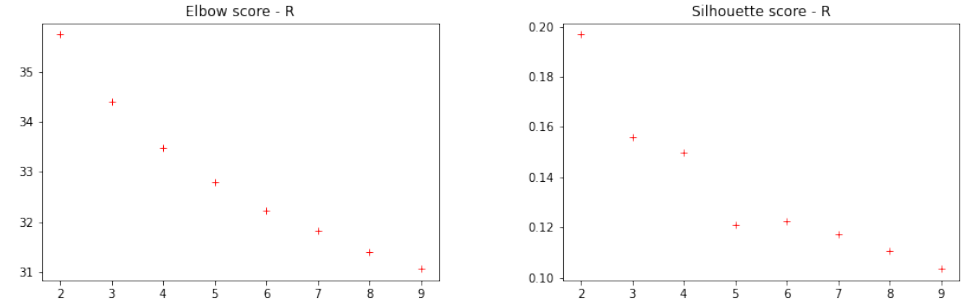

Figure 3.6 – Silhouette and elbow score - R

We present the reorganisation of the distance matrices according to the first 4 values of k (figure 3.7 for the left hemisphere and figure 3.8 for the right hemisphere).

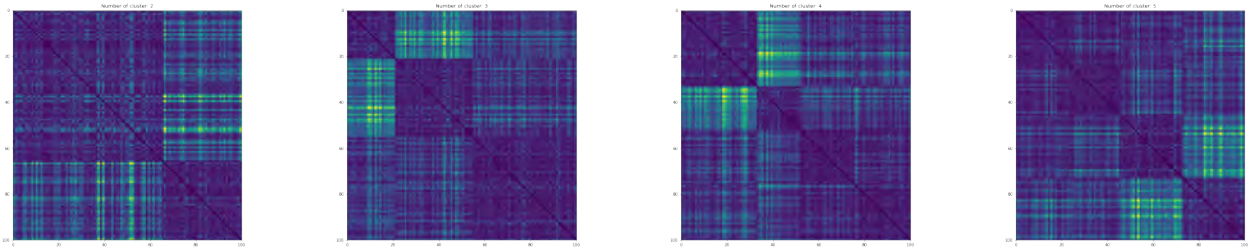

Figure 3.7 – Reorganised W_2^2 matrices - L

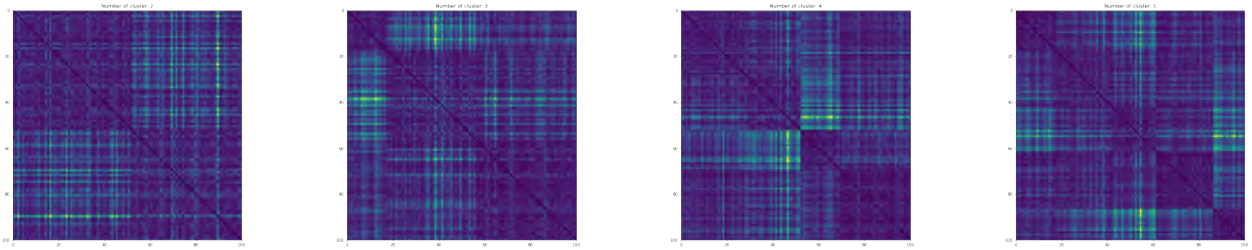

Figure 3.8 – Reorganised W_2^2 matrices - R

There seems to be a bend in the elbow score for k = 2 or k = 3 and that the silhouette score indicates k = 2 (figure 3.5 for the left hemisphere and figure 3.6 for the right hemisphere). We are interested in the study of these two cluster values.

Experiment 2: Barycenter of subgroups

We are working on establishing the barycenters for the subgroups for k = 2 and k = 3 since we cannot decide between these two numbers of clusters. For that, one determines the local maximum for each barycenter of a subgroup, for the two hemispheres.

hemisphere	left		right	
Cluster	0	1	0	1
Ventral	[6 5]	[5 5]	[6 5]	[5 5]
		[13 18]		
		[24 19]		[23 22]
		[31 30]		
	[67 66]	[63 64]	[67 67]	[66 64]
	[74 74]	[74 72]	[73 75]	[68 69]
	[80 83]	[87 90]		[80 84]
Dorsal	[95 97]	[95 96]	[96 97]	[95 97]

Table 3.5 – Position of local maximum - 2 clusters

hemisphere		left			right	
Cluster	0	1	2	0	1	2
Ventral	[6 5]	[5 5]	[7 6]		[5 5]	[6 5]
		[24 19]	[17 19]		[25 25]	[18 16]
		[31 29]				
				[51 58]	[43 49]	
	[69 69]	[63 64]	[66 64]	[67 68]	[65 62]	[67 66]
		[74 72]	[73 74]			[76 78]
		[87 91]		[79 86]	[79 83]	
Dorsal	[96 97]	[95 96]	[94 96]	[96 97]	[95 97]	[96 97]

Table 3.6 – Position of local maximum - 3 clusters - both hemispheres

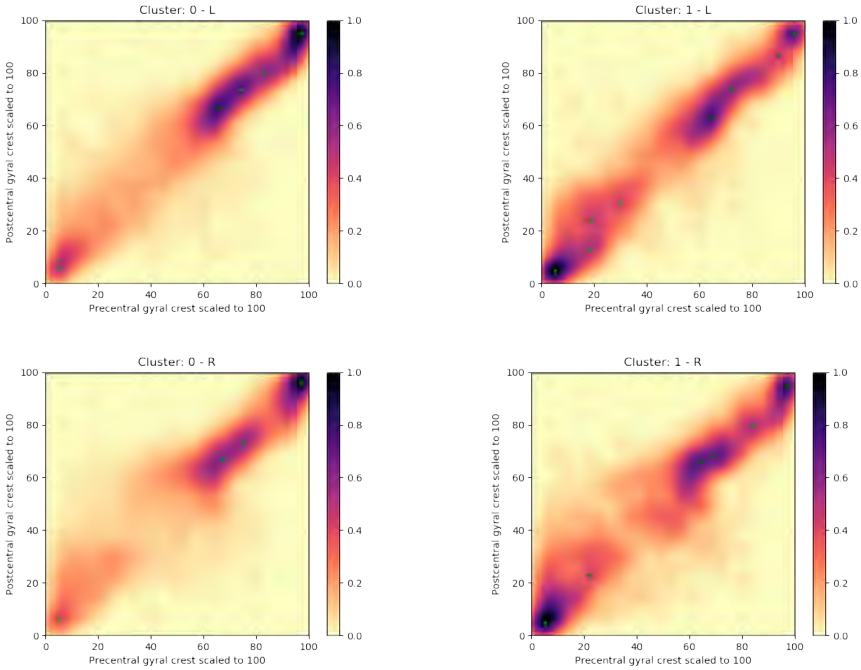

Figure 3.9 – Sub-cluster barycenter k = 2 - both hemispheres

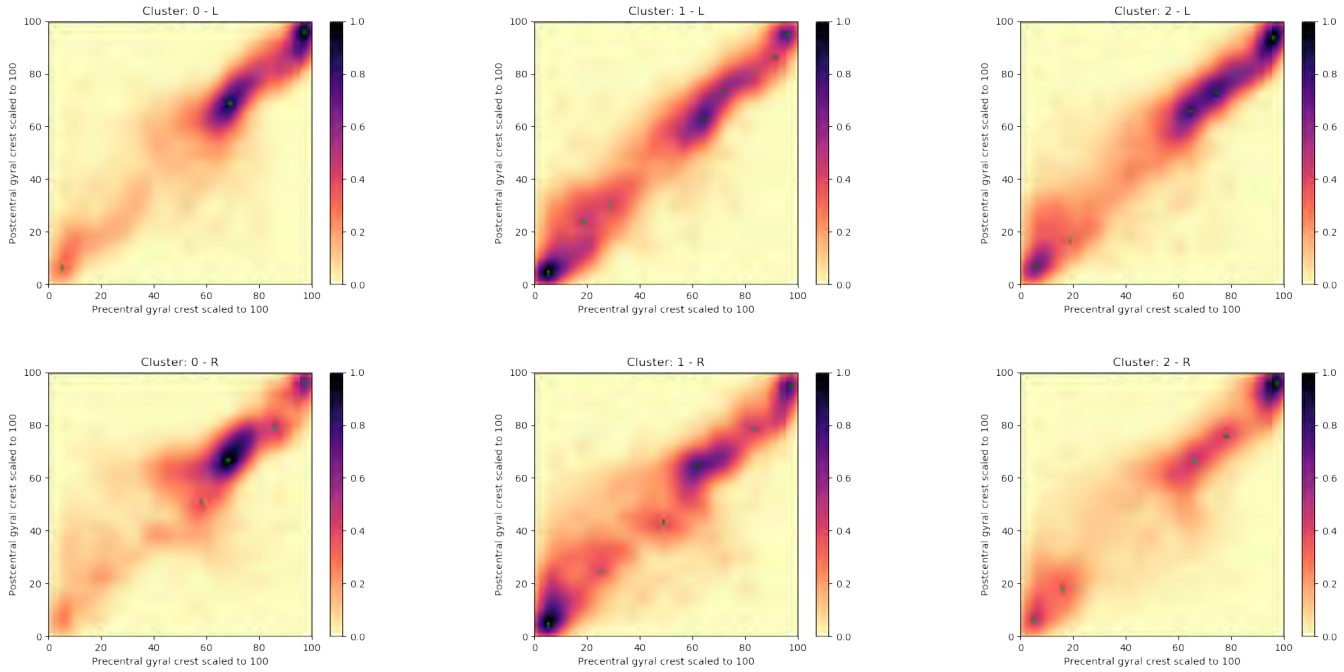

Figure 3.10 – Sub-cluster barycenter k = 3 - both hemispheres

These barycenters of subgroups allow us to see that we have distinct characteristics between the different clusters (k = 2 figure 3.9 and table 3.5, k = 3 figure 3.10 and table 3.6). We observe that there are points, remaining constant from one cluster to another and from one grouping to another, notably for the extremum points. As a reminder, the one on the bottom left corresponds to the mouth and the one on the top right to the foot.

3.2.2 Isomap

We project our data into a smaller dimensional space in order to carry out a qualitative analysis of the structures within our subjects. To do this, we use the isomap algorithm present in the manifold library of sklearn. We provide the 2-Wasserstein distance matrix, the desired dimension and the number of neighbours to consider in order to build our graph.

We try to define the neighbourhood hyperparameter in an optimal way. We present the different projections according to the number of neighbours. We associate the labels obtained during the previous experiment for k=2 and k=3. For the following part, n is the dimension of the isomap. Figures without labels are available in the appendix for each hemisphere and dimension of the isomap (n=2 figure 3.26 and 3.27 , n=1 figure 3.28 and 3.29).

3.2.3 Experiment 1: Isomap in dimension 2

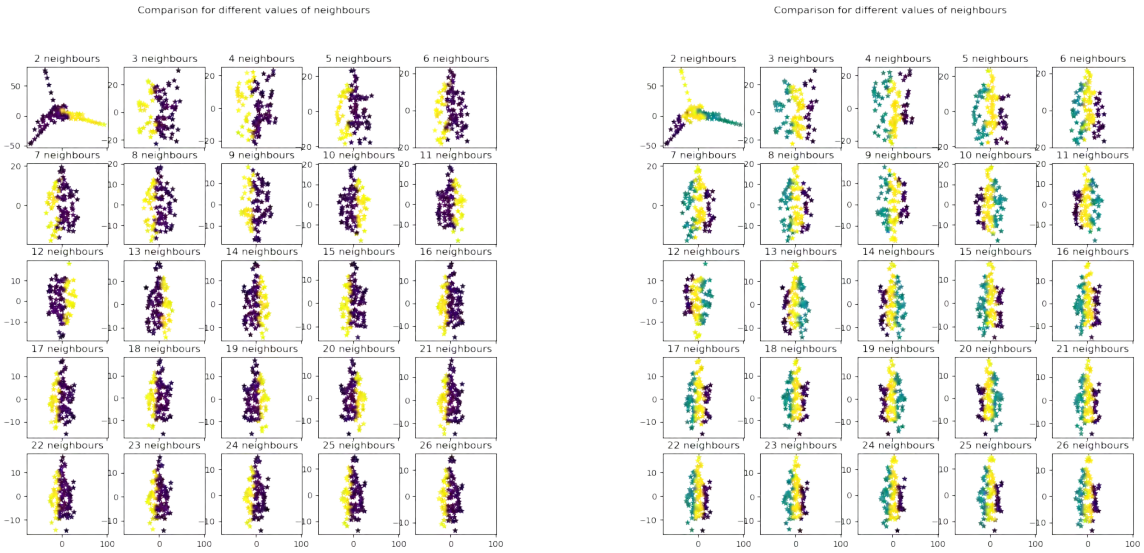

Figure 3.11 – Isomap dim 2 - L

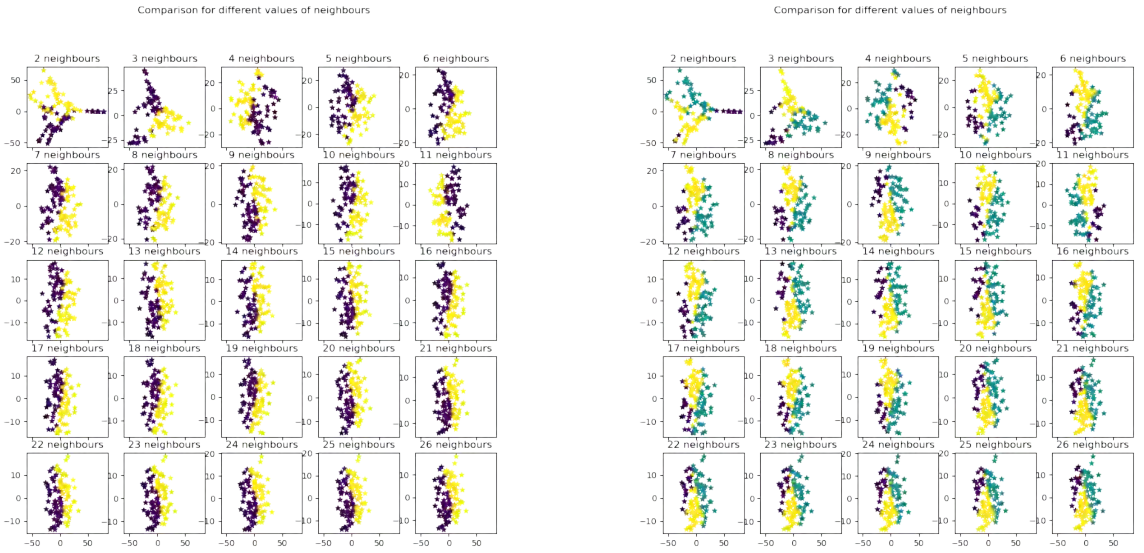

Figure 3.12 – Isomap dim 2 - R

It is difficult to choose an optimal value for the number of neighbours. At this stage, we could not make a rigorous implementation in order to determine this optimal value by the method proposed in *Selection of the Optimal Parameter Value for the Isomap Algorithm*[SMR06]. This method consists in summing the distances of all the bridges constituting the isomap, and to determine an inflection point of the curve of the previous score according to the number of neighbours.

Qualitatively, we know that the greater the number of neighbours, the more compact the distribution of subjects is. This explains the evolution of the figures. And we can use this

information to qualitatively determine the variation in compactness between figures. In this way we can determine that for a number of neighbours equal to 7, in dimension 1 and 2, we should have the elbow at the level of the curve because the figures evolve little around this neighbourhood. It is interesting to observe that in the case n=2, the labels provided by the K-medoids are successively juxtaposed along the x-axis whatever the value of k (figure 3.11 and 3.12).

3.2.4 Experiment 2: Isomap in dimension 1

Figure 3.13 – Isomap dim 1 - L

Figure 3.14 – Isomap dim 1 - R

With the results of experiment 1, we can see that the isomap n=1 is the result of the projection on the x-axis of the isomap n=2. The same organisation of the subjects is observed, albeit

with opposite coordinates (figure 3.13 and 3.14). The order of the subjects does not change in the case n=1. We are interested in an axis of variations. In this case, the incidence of the number of neighbours is not very important.

We observe that the clusters determined by the K-medoids are quite distinct for the values k=2 and k=3, for the values above we have more confusion. These observations encourage us in the fact that we can determine subgroups of populations. It could be that these subgroups do not correspond to those determined by the K-medoids if we do a K-means in the space generated by the isomap, especially for n=2. Insofar as the geodesic distances are kept, the most similar subjects are supposed to be the closest in space, However, we do not observe this distribution in 2 dimensions, especially for k=2. It would seem that this is more the case with k=3.

3.2.5 Experiment 3: Sliding Barycenter

Taking into account the preceding experiments, we choose to study the possibility of having a continuous distribution within the population of a characteristic. For this, we define an axis of variation provided by the isomap n=1. This hypothesis is based on the article *The effect of handedness on the shape of the central sulcus* [Sun+12] it was highlighted that there is an axis of variation depending on the position of the hand knob in the central sulcus (figure 3.15). We expect to have an equivalent result.

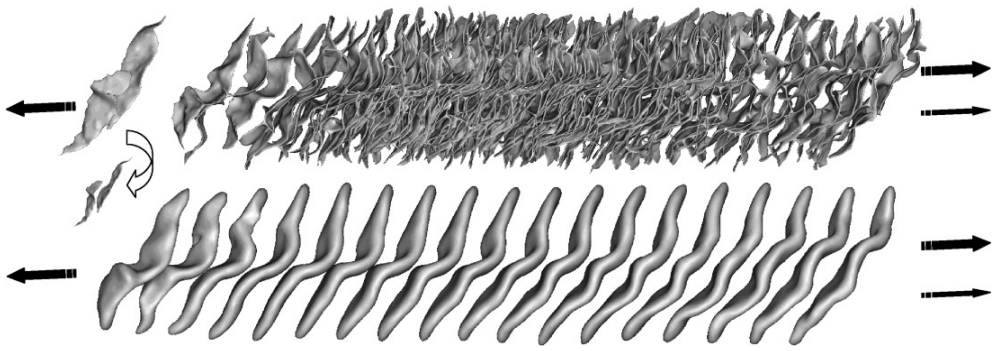

Figure 3.15 – Hand knob

For this purpose, the sliding barycentre is calculated with different window sizes in order to highlight an evolution between the different subjects along this axis of variation. We look for the smallest possible window allowing us to follow the maxima in a relevant way. If the window is too small, we have a sliding barycenter that fluctuates too much from one subject to another, and if the window is too large, we lose the notion of variation from one barycenter to another.

We define empirically the size of the window at 15 subjects before and 15 subjects after the one concerned, i.e. a sliding window of 31 subjects out of our 100. The local maxima for each barycenter are determined in such a way that the position of the central task corresponding to the hand can be followed. We present a sample of the barycenters generated in figure 3.16 and 3.17, and a graph summarising the position of the local maxima of the central zone (figure 3.18).

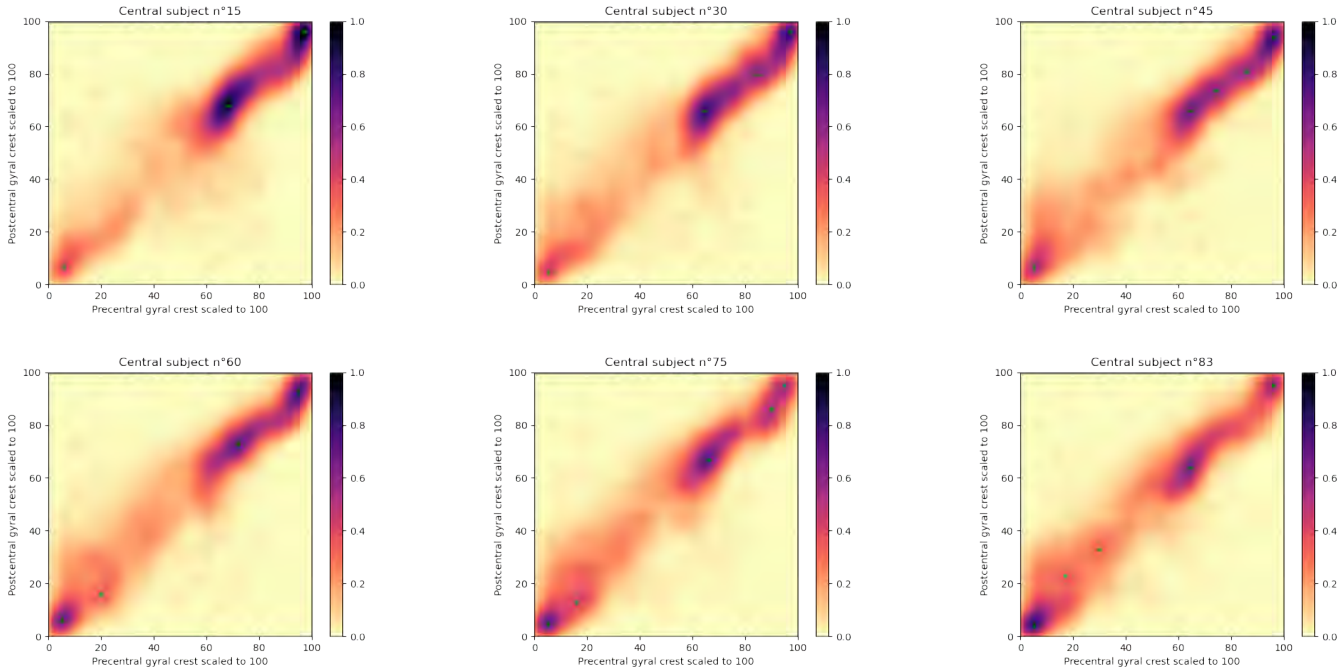

Figure 3.16 – Samples of sliding barycenter - L

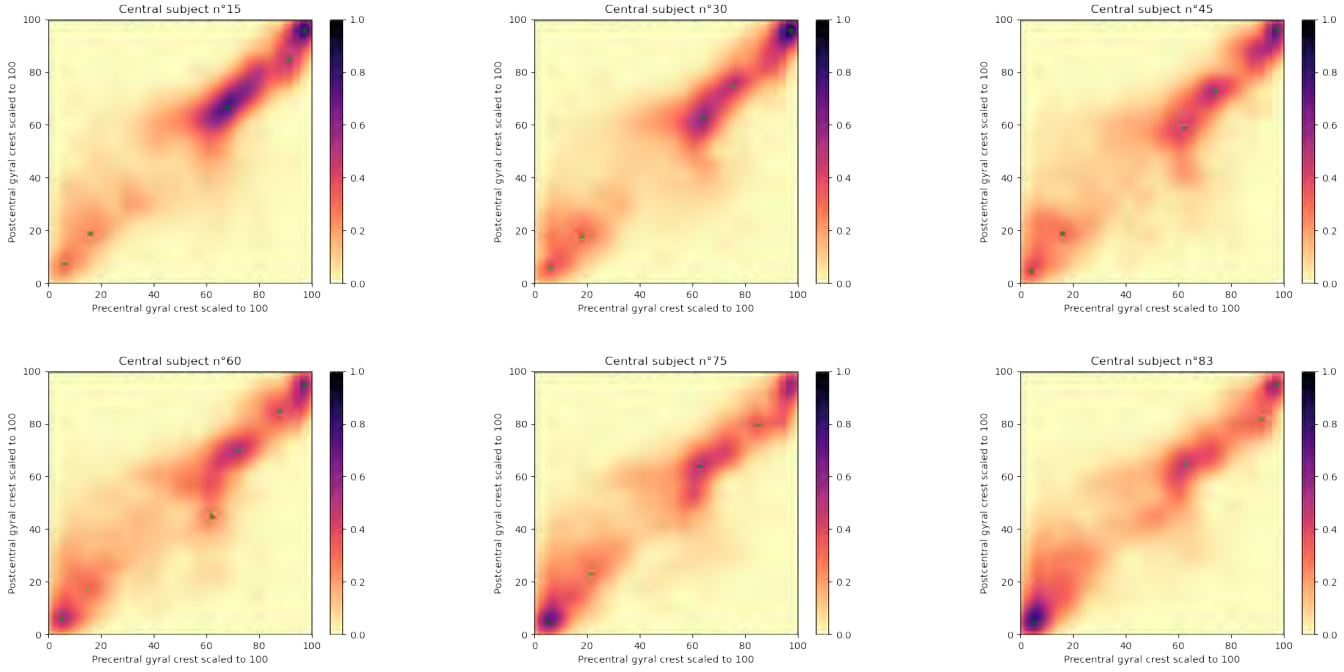

Figure 3.17 – Samples of sliding barycenter - R

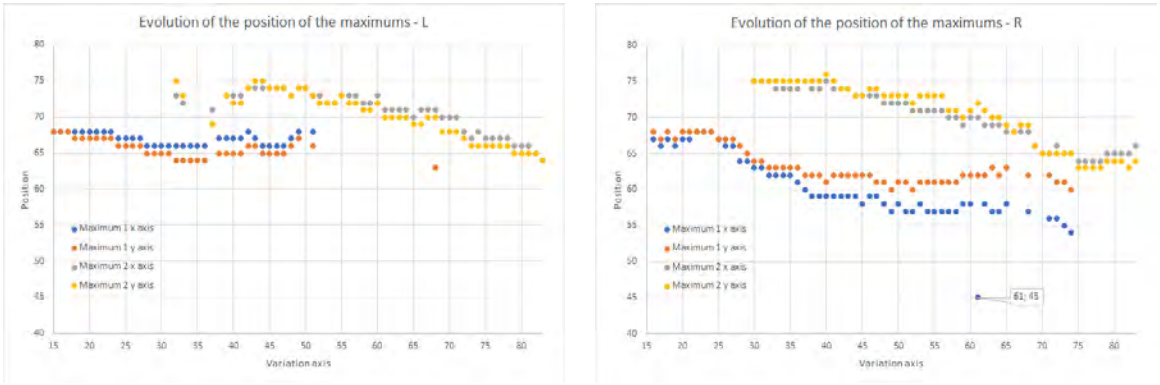

Figure 3.18 – Position of the maximums of the central area

A strong correlation between the position of the central task and the axis of variation provided by the isomap can be seen in figure 3.18. It seems that there is also a correlation with the variance of the task according to this axis of variation. These results can be put in perspective with those obtained in *The effect of handedness on the shape of the central sulcus* [Sun+12]. It seems that we have the same observation with the connectivity maps as in the previous study with the position of the hand knob. In our case we would observe a shift in the position of the connectivity density corresponding to that of the hand knob.

3.2.6 Conclusion

Through this study, we see that the notion of clusters is not necessarily the most relevant in our case. In fact, we have a continuous distribution of subjects according to the position of the central zone corresponding to the hand. We can study the characteristics of the sliding barycenters in more detail to better highlight the differences. These results can be put in perspective with another study related to the hand knob, which is an anatomical bitter of the central sulcus.

3.3 Correspondence between group subjects and individual subjects

3.3.1 Gaussian mix to Gaussian cloud

In this experiment we try to understand the behavior of the optimal transport in the case where one determines the matrix of transport of a Gaussian mixture towards a Gaussian cloud, in the case discrete as continuous.

Discrete case

We have chosen to label the points in this case according to their coordinates in order to have a better representation after transport. We simulate a point cloud (array 3.7).

Simulation	Nature	Points	Algorithm
Source	Gaussian Gs = G1 + G2	Gs: 400,700	EMD
Target	Gaussian Gt	Gt: 1000	

Table 3.7 – Gaussian mixture to Gaussian cloud - discrete case

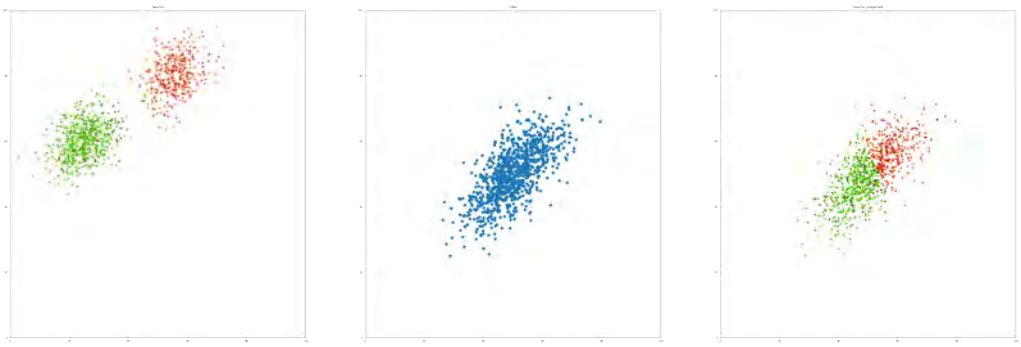

Figure 3.19 – Gaussian mixture to Gaussian cloud

We note that we distinguish on the projected source the origin of the source points corresponding to each cloud (figure 3.19). This transport corresponds to our expectations.

Continuous case

As a reminder, we generate the continuous case from the discrete case. We made the choice to label according to the amplitude of the values in the continuous case.

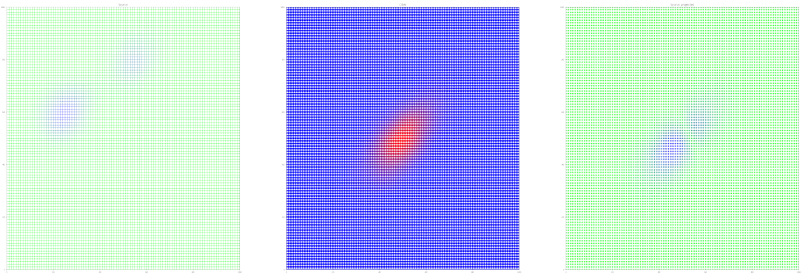

Figure 3.20 – Gaussian mixture to Gaussian cloud - continuous case

We observe in figure 3.20a difference compared to the discrete cases 3.19in particular on the separation of the two original Gaussian clouds. This is explained by the fact that to generate the continuous mapping we have to convolve a Gaussian kernel which induces that we do not have strictly zero values on the density mapping, even for points very distant from the point cloud. In addition, the value associated with our labels also contributes to the illusion of an existing border between the two original Gaussians.

In order to solve these problems, we can do a threshold to set irrelevant values to 0. This has the advantage of reducing the computation time as well. However, this is not a choice that we made since it would add a hyperparameter to our problem. Or we can also rework the color scale of the labels to reduce this unwanted border effect.

3.3.2 Entropy regularisation - discrete mappings

We want to know the impact of entropy regularisation on the solution of the optimal transport problem. For that one uses the algorithm of Sinkhorn for various values of regularisation. We also compare the optimal transport matrices resulting from the projection established from them.

We keep the characteristics of the point clouds generated in the table 3.7. One changes the algorithm to use the algorithm of Sinkhorn with the values of regularisations between 0.001 and 100 incremented in power of 10. One presents 3 cases on the 6 which we made, that is to say the values of following regularisations: 0.01, 0.1, 1.

Optimal transport matrix

Figure 3.21 – OT matrix - entropic regularisation

Projection

We have chosen to display the link between the source point projected in the target space with the original source point, in addition to the labels in the figure 3.22. This makes it possible to mount the effect of the regularisation on the projection of the source on the target. Remember that we only display the contribution of the most important point both for the label and for the connection, that is, the one with the greatest weight in the optimal transport matrix (figure 3.21).

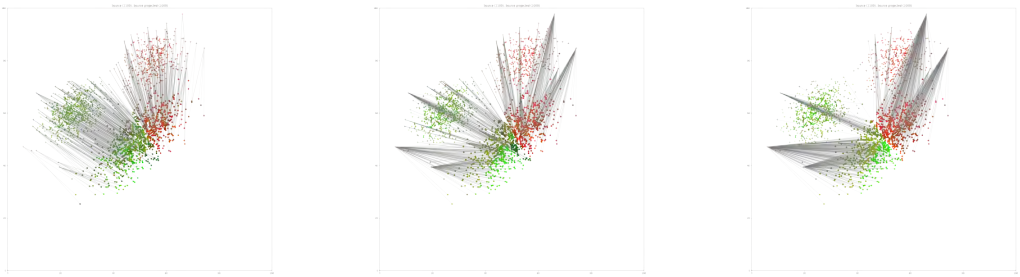

Figure 3.22 – Source projected - entropic regularisation - discrete case

Conclusion

It is noted that for weak values of regularisations, less than 0.1, and considering the predominant weight as the value defining the label, the results are close to the non-regularised solution, the contribution of a source point is little spread compared to the higher values of regularisation. There is a strong resemblance between the transport matrices in terms of texture. This can be explained by the fact that we can see entropy regularisation as a constrained resolution with margins.

Insofar as we have an approximation of the optimal transport matrix without regularisation, we do not have strictly identical results between the transport from the source to the

target, and from the target to the source, except for a transpose, even if there is a strong resemblance. This is information to take into account in the case where one seeks to make the reverse return from the target to the source, in particular to back-propagate information.

3.3.3 Entropy regularisation - continuous mappings

We reproduce the same experiments as in the previous experiment but with continuous mappings, derived from discrete data.

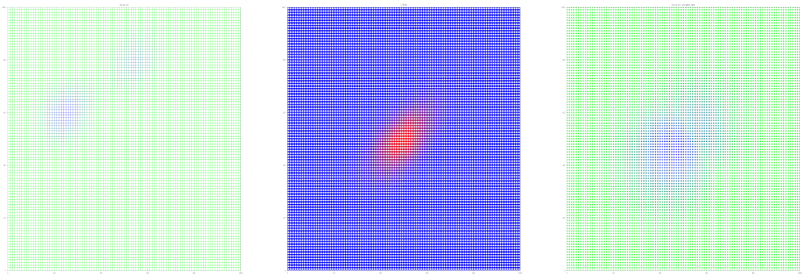

Figure 3.23 – Source projected - entropic regularisation (reg = 100) - continuous case

We observe the same phenomenon as before on the spreading of the weights as the regularisation is large, and this results in the projection by a more important spread of the labels associated with the high values in their neighborhood (figure 3.23). However, we keep the border delimiting the two original clouds. In this case, we have a similar transport between the discrete mapping and the continuous mapping.

In addition, we cannot apply the regularisation with lower values in our case and that for large values of entropy regularisation we do not save time, quite the contrary. As indicated in the POT library, we can have a faster solution with the EMD algorithm. This has the advantage of providing us with a value close to the real transport matrix.

3.3.4 Simulation of real data

We deepen the investigations that we made previously by using more complex simulations. From the thesis of Pron, we get the results of the clusters and their characterisation in order to artificially generate the ideal group representation.

Real data simulation

Anatomical group	Nature	Points	Position
G1			
Name	$cov = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$	120	$\begin{bmatrix} 9 \\ 9 \end{bmatrix}$
Main			
G1	$cov = \begin{bmatrix} 2 & 0 \\ 0 & 10 \end{bmatrix}$	40	$\begin{bmatrix} 55 \\ 50 \end{bmatrix}$
G2	$cov = \begin{bmatrix} 5 & 5 \\ 5 & 15 \end{bmatrix}$	120	$\begin{bmatrix} 65 \\ 65 \end{bmatrix}$
G3	$cov = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$	80	$\begin{bmatrix} 75 \\ 75 \end{bmatrix}$
G1			
Name	$cov = \begin{bmatrix} 2 & 0 \\ 0 & 20 \end{bmatrix}$	120	$\begin{bmatrix} 95 \\ 90 \end{bmatrix}$

Table 3.8 – Simulation of real data

From the simulation we made (table 3.8), we introduce parasitic sets of Gaussian nature. The goal is to generate a large number of different subjects from the same database. We seek to observe the effects of parasitic sets on optimal transport. We choose the number of parasitic nuclei and the proportion thereof relative to the ideal group representation. It is assumed that the parasitic sets are in the minority compared to the other sets.

Optimal transport without regularisation and with entropy regularisation

We study the effect of noise on the optimal transport and the projection of the source on the target space, with or without regularisation.

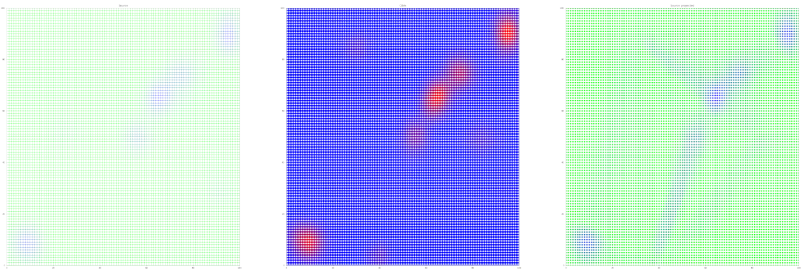

Figure 3.24 – Entropic regularisation (reg = 10) - continuous case

We have no perspective difference at the end of the labeling of the source in the target space, with or without entropy regularisation (figure 3.24). We see that there is a spread of very important values because of the noises that we have generated. In some simulations these noises have a low impact, but in the case where these are very far from the sets of interests we have a deformation of the Gaussian clouds.

This is the same phenomenon that we could see in the experiment of the two Gaussian clouds towards only one: this is due to the fact that we do not have strictly zero values for

points very far from Gaussian clouds, and that this increases the spreading of values when matching against parasitic sets.

Optimal transport with entropy and Tikhonov regularisation

We try to solve the previous problem by adding a regularisation term acting on the amplitude of the values of the optimal transport matrix. It is a generalisation of the Tikhonov problem adapted to the 2-Wasserstein distance.

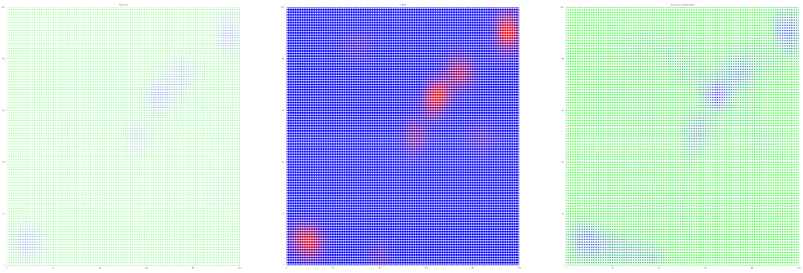

Figure 3.25 – Entropic and Tikhonov regularisation (reg1 = 10, reg2 = 10) - continuous case

We observe an improvement of the results (figure 3.25). There is a less important spread of the values thanks to this regularisation on the amplitudes of the values. However the result is far from perfect and in the case of real data there is a risk of having more parasitic sets.

3.3.5 Conclusion

Through this preliminary work, we choose to work on discrete connectivity mappings. This choice is motivated by the calculation time but also the quality of the transport. As we have been able to highlight, the generation of continuous mappings presents a significant drawback due to the way in which they are generated: we have the presence of quasi-null values but not zero. However, these values are also subject to the projection during the calculation of the optimal transport. and that creates transfer artifacts for us. Or by working directly on continuous mappings at each stage of the process and by applying the process to obtain a continuous connectivity mapping, we have very good results.

However, if we want to continue our work on these continuous mappings, a more detailed investigation of the methods of regulation should be made. One proposal we made is to act on the amplitude of the values in order to minimise the spread of the latter with the generalised Tikhonov regularisation at the 2-Wasserstein distance.

Conclusion

The aim of our work has been to use the tools of optimal transport to improve existing results and to improve existing results and to explore new avenues. In this respect, we have succeeded in generating a more representative group subject. For this, we used the 2-Wasserstein metric to evaluate this improvement. We also highlighted the most central topic in the set of our topics.

Following this, we sought to characterise our topics in the sense that we sought to study the relationships between them. That is to say, if they were grouped in sub-groups or if on the contrary we have a continuous evolution of a characteristic that we have to determine. It would seem that at the end of our investigation that if there is the existence of a sub-group it is less than or equal to 3. A more relevant result is the continuous variation of the position of the central spot corresponding to the hand. This result was obtained by classifying the subjects on a variation axis and We then studied the result from a sliding barycenter on all our subjects along this axis. Further investigation should be made to go back to the 3D space of the connectivity map to observe these variations.

One of the tasks we had was to improve the mapping between a group subject and the individual subjects. This topic was initiated but not pursued to completion. It is indeed possible to match two subjects in order to retropagate, in particular, segmentation labels on the group subject to the individual subjects but we have not set up a protocol. It would be interesting to study in more detail the influence of different regularisations and to apply them on real cases.

More generally, it would be interesting to make a comparative study between the left and right hemispheres in order to observe if there are similarities, especially for the axis of variation. All of our experiments and results are freely available and can be reproduced.

Appendices

3.4 Centroid

Name	ΣW_2^2	$\overline{W_2}$	Number of points	Name	ΣW_2^2	$\overline{W_2}$	Number of points
495255_L	20508.05	14.39	2639	156637_L	29651.87	17.31	1849
598568_L	20752.38	14.48	1825	114419_L	29653.65	17.31	3668
118124_L	20799.91	14.49	2509	123117_L	29817.50	17.35	2549
204420_L	20902.14	14.53	1898	268850_L	30854.74	17.65	4468
147030_L	21242.92	14.65	1959	432332_L	30871.53	17.66	2090
759869_L	21253.39	14.65	3038	152831_L	31463.00	17.83	3106
583858_L	21491.06	14.73	1639	201111_L	31674.92	17.89	1967
156233_L	21552.21	14.75	2941	144428_L	31717.59	17.90	2861
845458_L	21536.32	14.75	3756	103414_L	31803.39	17.92	2313
173536_L	21777.56	14.83	2406	303119_L	32021.67	17.98	3223
198350_L	22079.23	14.93	2918	103111_L	33150.92	18.30	2173
154936_L	22392.52	15.04	2745	100206_L	33488.92	18.39	2029
773257_L	22600.21	15.11	1336	153429_L	33835.72	18.49	2354
188448_L	22893.73	15.21	1550	867468_L	34268.25	18.60	2669
207123_L	22965.95	15.23	1877	456346_L	34295.49	18.61	2041
173940_L	23187.50	15.30	2605	211417_L	34569.71	18.69	2175
346137_L	23214.50	15.31	1842	381543_L	34642.31	18.71	2526
108828_L	23366.08	15.36	3131	180129_L	34785.95	18.74	2032
802844_L	23388.39	15.37	3607	992673_L	35347.73	18.90	2798
942658_L	23511.79	15.41	1448	445543_L	36092.59	19.09	1450
158540_L	23534.70	15.42	2575	186444_L	37021.88	19.34	2280
134425_L	23545.43	15.42	2696	214423_L	37238.29	19.39	2189
173637_L	23798.90	15.50	2151	117324_L	37277.07	19.40	3311
157942_L	23836.46	15.52	1865	144832_L	37527.14	19.47	2016
121618_L	24172.92	15.63	3376	192136_L	37879.46	19.56	1697
118730_L	24503.99	15.73	2413	540436_L	37869.37	19.56	3622
188347_L	24819.97	15.83	3520	673455_L	38422.73	19.70	2677
195950_L	24926.52	15.87	2295	196346_L	38579.99	19.74	2198
545345_L	24920.60	15.87	3432	947668_L	39618.45	20.00	1423
171532_L	25133.85	15.93	3553	788876_L	39597.70	20.00	3000
210415_L	25463.20	16.04	1866	214726_L	39684.72	20.02	1478
132017_L	25641.02	16.09	1637	212116_L	40520.45	20.23	3012
304020_L	26026.13	16.21	3857	285345_L	40886.89	20.32	3381
882161_L	26104.00	16.24	2977	117930_L	41354.45	20.44	2239
371843_L	26439.55	16.34	2559	189450_L	41471.10	20.47	1565
191942_L	26469.86	16.35	2153	140319_L	41974.74	20.59	2456
136732_L	26668.17	16.41	3176	580650_L	42476.12	20.71	1852
346945_L	27197.53	16.57	943	156031_L	42847.48	20.80	2059
194645_L	27274.62	16.60	3148	657659_L	44026.95	21.09	2086
843151_L	27300.56	16.61	2592	150726_L	45732.47	21.49	2886
212217_L	27919.61	16.79	2176	810843_L	47636.34	21.94	2967
580751_L	28193.05	16.88	2166	107321_L	48125.99	22.05	3085
193239_L	28470.16	16.96	2870	599065_L	54004.50	23.36	1927
138231_L	28527.17	16.98	4413	194746_L	56757.61	23.94	2989
530635_L	28624.60	17.00	1759	144125_L	59335.18	24.48	1641
211316_L	28659.37	17.01	1701	119126_L	60384.91	24.70	2333
108525_L	29077.58	17.14	2695	129129_L	62401.77	25.11	2984
114621_L	29308.23	17.21	2294	130821_L	66839.09	25.98	1759
571144_L	29365.82	17.22	3041	580044_L	67501.00	26.11	2003
555651_L	29424.57	17.24	3491	121921_L	68915.71	26.38	3382

Table 3.9 – Centroid - distance of one subject from the other 100 - left hemisphere

Name	ΣW_2^2	$\overline{W_2}$	Number of points	Name	ΣW_2^2	$\overline{W_2}$	Number of points
123117_R	22110.10	14.94	3190	156637_R	32293.03	18.06	1499
545345_R	22337.42	15.02	2832	802844_R	32644.86	18.16	1694
118730_R	22406.80	15.04	2559	201111_R	32802.67	18.20	1776
114621_R	23374.32	15.37	2412	147030_R	33527.96	18.40	1908
212217_R	23482.13	15.40	2159	346945_R	33759.76	18.47	1557
191942_R	23930.52	15.55	1634	580751_R	33995.52	18.53	2691
788876_R	24066.98	15.59	2374	381543_R	34220.93	18.59	1757
121921_R	24224.67	15.64	2296	992673_R	34263.63	18.60	2786
144428_R	24449.76	15.72	4029	207123_R	34400.74	18.64	1819
673455_R	25233.60	15.97	2376	432332_R	35354.50	18.90	1920
134425_R	25349.94	16.00	1757	157942_R	35398.93	18.91	1259
192136_R	25352.41	16.00	2708	129129_R	35565.16	18.95	1362
173637_R	25676.67	16.10	2232	144832_R	35598.10	18.96	1665
108828_R	25759.28	16.13	3083	495255_R	35918.02	19.05	3795
194645_R	25947.41	16.19	3090	130821_R	36294.35	19.15	1820
121618_R	26061.88	16.23	2040	188347_R	36593.74	19.23	2508
152831_R	26781.34	16.45	2259	285345_R	37072.52	19.35	2697
117930_R	26911.24	16.49	1528	150726_R	37201.69	19.38	2568
657659_R	27067.06	16.53	1343	194746_R	37411.36	19.44	1977
171532_R	27209.38	16.58	2074	156031_R	37458.25	19.45	2104
210415_R	27279.87	16.60	1764	211316_R	38109.53	19.62	1541
346137_R	27487.56	16.66	1778	154936_R	38129.45	19.63	2042
845458_R	27517.36	16.67	2578	196346_R	38882.78	19.82	1865
304020_R	27806.62	16.76	1798	144125_R	39800.63	20.05	1168
204420_R	27861.52	16.78	1555	198350_R	39993.60	20.10	3473
882161_R	27934.76	16.80	2283	810843_R	40035.25	20.11	2066
173536_R	28078.46	16.84	2587	118124_R	40365.93	20.19	2823
186444_R	28130.35	16.86	1532	158540_R	41121.03	20.38	2151
530635_R	28214.28	16.88	2307	867468_R	42459.34	20.71	1454
156233_R	28250.34	16.89	2279	136732_R	42487.65	20.72	2429
214423_R	28314.95	16.91	2707	132017_R	45474.53	21.43	1297
214726_R	28711.08	17.03	1514	555651_R	45843.12	21.52	1910
540436_R	28697.39	17.03	2795	942658_R	46101.84	21.58	964
114419_R	28767.29	17.05	2719	268850_R	46866.68	21.76	3613
108525_R	28802.36	17.06	2404	580650_R	47578.52	21.92	1052
371843_R	28847.06	17.07	1789	773257_R	47976.77	22.01	1579
598568_R	28967.02	17.11	1748	103414_R	49500.98	22.36	1938
173940_R	29580.38	17.29	1925	107321_R	50119.07	22.50	2236
211417_R	29744.93	17.33	1788	119126_R	52433.52	23.01	1206
140319_R	30034.90	17.42	1202	193239_R	53496.83	23.25	2665
456346_R	30040.83	17.42	2198	599065_R	56281.97	23.84	768
843151_R	30230.00	17.47	2786	583858_R	57349.74	24.07	2236
195950_R	30264.86	17.48	1394	103111_R	57683.26	24.14	2079
759869_R	30289.69	17.49	1798	138231_R	58092.12	24.22	3088
180129_R	30618.10	17.59	1238	117324_R	61298.79	24.88	2267
189450_R	31701.25	17.89	1742	580044_R	61503.30	24.92	1472
571144_R	31858.50	17.94	1665	212116_R	62795.58	25.19	2911
153429_R	32070.57	18.00	2654	947668_R	73683.96	27.28	1541
445543_R	32161.40	18.02	1333	100206_R	76526.24	27.80	1878
303119_R	32152.92	18.02	3413	188448_R	100505.67	31.86	582

Table 3.10 – Centroid - distance of one subject from the other 100 - right hemisphere

3.5 Isomap

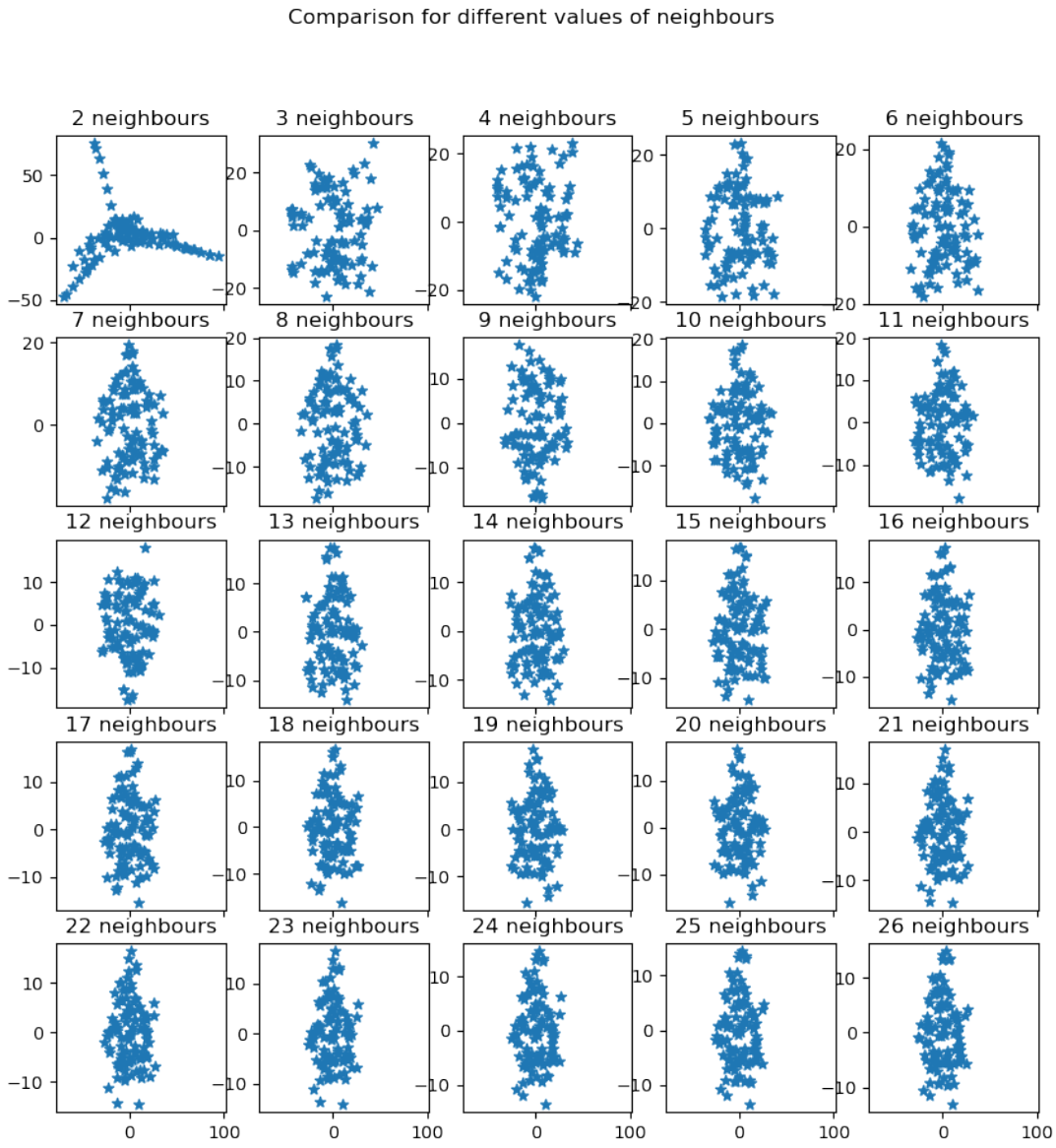

Figure 3.26 – Isomap dim 2 - L

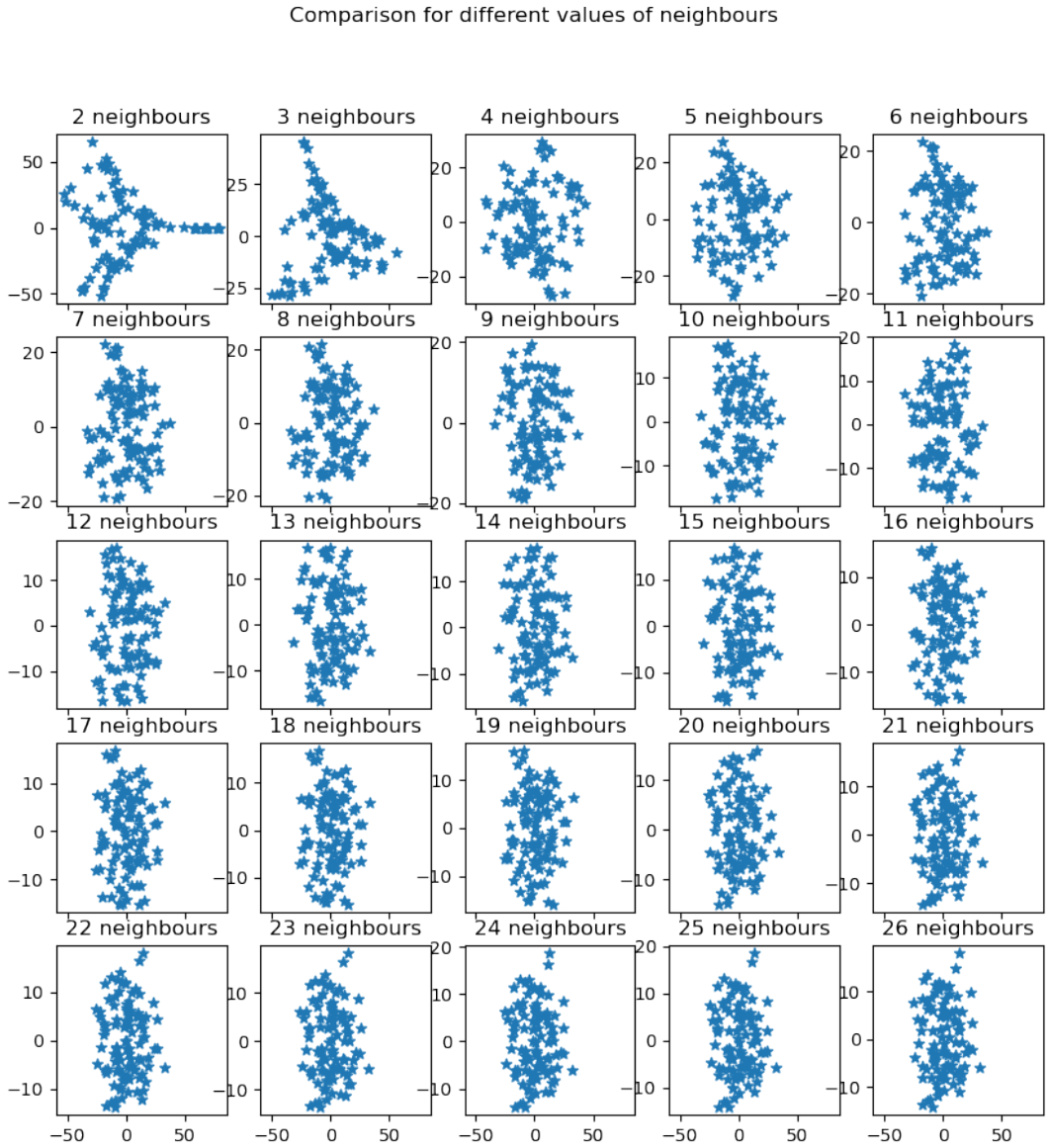

Figure 3.27 – Isomap dim 2 - R

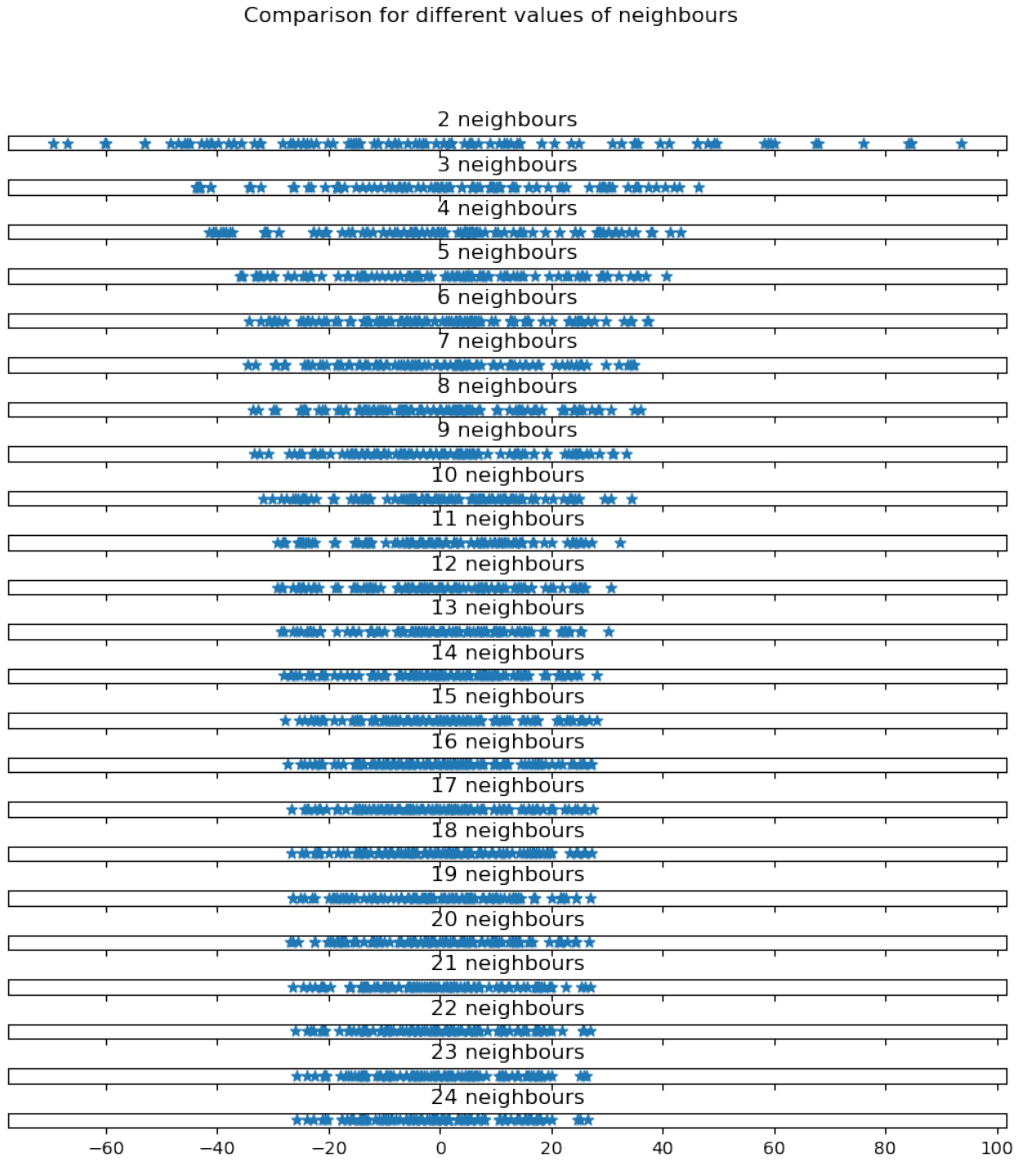

Figure 3.28 – Isomap dim 1 - L

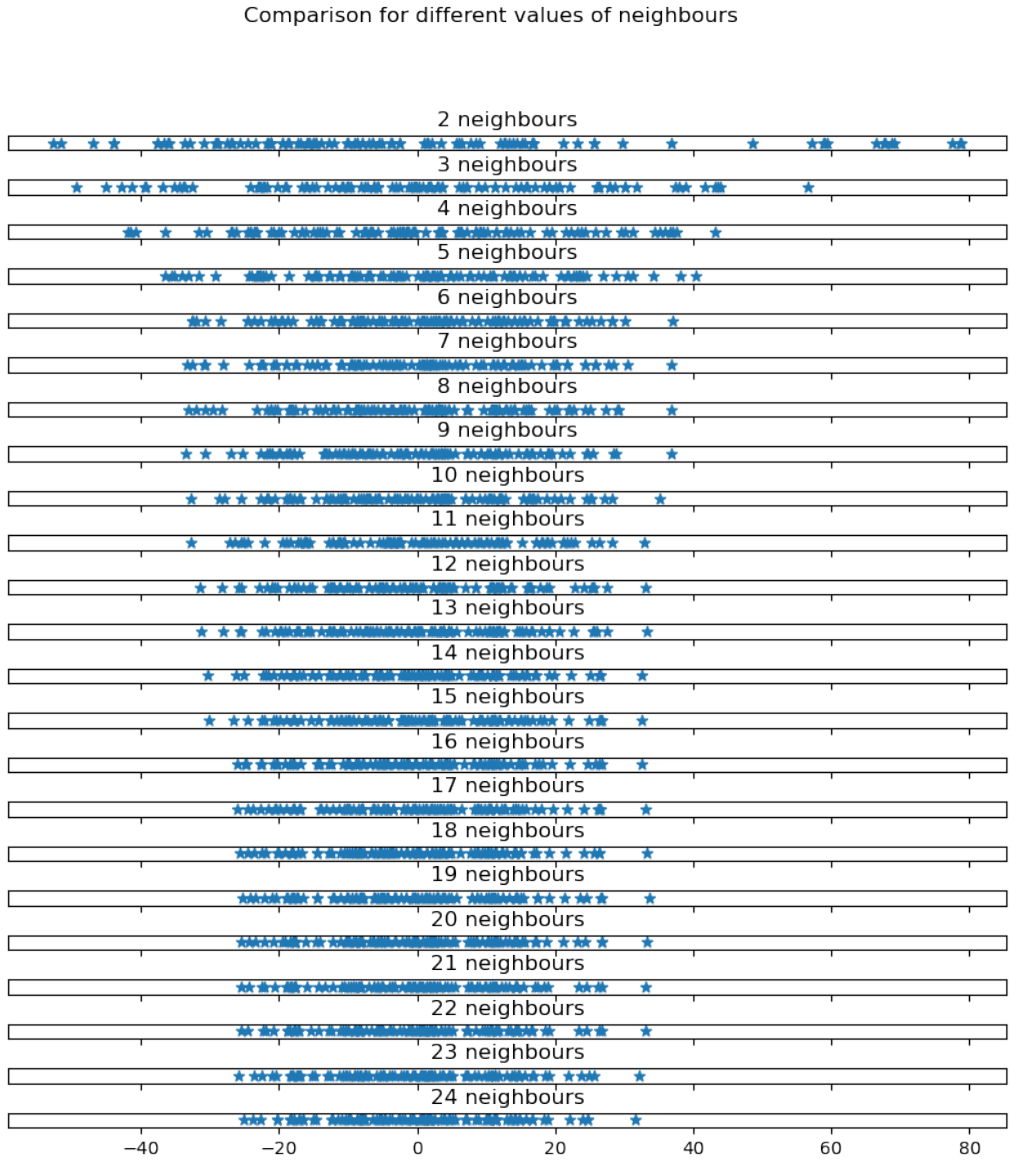

Figure 3.29 – Isomap dim 1 - R