Morphotypes of the human body
a clustering approach.

Steve de Rose

Supervised by :
Frédéric Bertrand (UTT)
Myriam Maumy Bertrand (UTT)
Philippe Meyer (UTT)

Academic year 2021-2022

Contents

1 Project's Presentation

This project is part of the Master in Scientific Computing and Mathematics for Information. One of the objectives of this master is to provide its students with advanced skills in data analysis. This work is an excellent opportunity to put into practice what has been learned throughout the master's degree.

1.1 DiTeX: Data-Innovation for the Textile Industry

DiTeX is a joint research and development laboratory between the University of Technology of Troyes and the French Institute of Textiles and Clothing.

This laboratory develops statistical modelling and machine learning to analyse data from clothing and to respond to problems dealing, in particular, with the measurements of the human body: What is the effect of ageing? What are the types of morphologies?

1.1.1 University of Technology of Troyes

Founded in 1994, the University of Technology of Troyes (UTT[1]) is a French university, in the Academy of Reims.

The UTT is part of the network of the three universities of technology, found by the University of Technology of Compiègne. Inspired by the American University of Pennsylvania in Philadelphia, these three universities (UTC, UTBM and UTT) are a French mixture between the universities of this country and its schools of engineers (Grandes Ecoles).

1.1.2 French Institute of Textiles and Clothing

The French Institute of Textiles and Clothing (IFTH[2]) is an industrial technical centre created by decree on 14 April 2001. Its mission is to promote and assist progress in the textile and clothing sectors.

1.2 Context

The variety of human morphologies is an important issue for the textile-apparel industry. Indeed, sizing systems currently used by companies have to be continuously updated or adapted to the population target.

For this reason, the Textile-Apparel-Industry requires a very accurate sizing system to minimize their costs and satisfy their customers. However, the specific constraints of human morphologies complicate the sizing system definition procedure and distributors prefer to use standard sizing system rather than an intelligent system suitable to their customers.

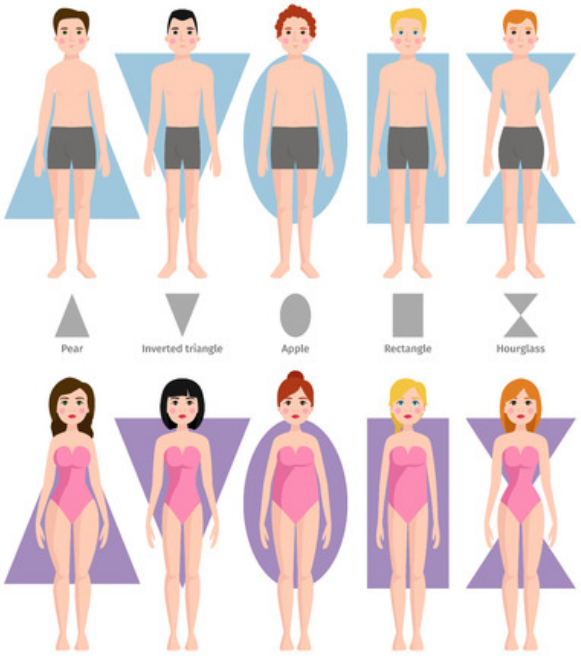
Until now, the morphotypes of a population are extracted from measurement charts. However, new technologies such as 3D body scanning open new opportunities to enhance

the morphotype generation from a sample of population especially with the 3D data of bodies.

The aim of this research is to define an exhaustive methodology to obtain a clustering of human morphology shapes representative of a population and to extract the most significant morphotype of each class. Clustering methods are implemented and the performances are evaluated using real data.

1.3 Classification of human body types

Body shapes are often categorised in the fashion industry into one of five elementary geometric shapes, though there are very wide ranges of actual sizes within each shape:



- **Triangle, "A" Frame, Pear, Spoon or Christmas Tree**
  The hips are wider than the bust. The distribution of fat varies, with fat tending to deposit first in the buttocks, hips, and thighs. As body fat percentage increases, an increasing proportion of body fat is distributed around the waist and upper abdomen.
- **Inverted Triangle, Cone or "V" Frame**
  The shoulders are broader than the hips. The legs and thighs tend to be slim, while the chest looks larger compared with the rest of the body. Fat is mainly distributed in the chest and face.
- **Oval, Circle/Rounded, Apple, Diamond or "O" Frame**
  Top and bottom are narrow. Chest and belly are where weight is found. Legs are skinny.
- **Rectangular, Ruler or "H" Frame**
  The waist is less than 23cm smaller than the hips and bust. Body fat is distributed predominantly in the abdomen, buttocks, chest, and face. This overall fat distribution creates the typical ruler (straight) shape.
- **Hourglass, Figure 8 or "X" Frame**
  The hips and bust are almost of equal size, and the waist is narrower than both. Body fat distribution tends to be around both the upper body and lower body. This body type enlarges the arms, chest, hips, and rear before other parts, such as the waist and upper abdomen.

The idea of classifying the morphotypes of the human body is ancient (Hippocrates in the 3rd century BC).

In 1941, the O'Brien and Shelton study of women's measurements[3] was one of the first to systematically collect linear body measurement data to be used for sizing apparel. No scientific study of body measurements used in the construction of women's clothing had ever been reported.[4]

In 1968, Douty and Brannon used face forward and side silhouette photographs to study female and male body build and posture. Their studies[5] advanced somatographic measurement methods and focused on the constructs of body build and posture by visually classifying body types. Douty's body build scales were derived through visual analysis of photographs projected onto a Cartesian grid structure for reference.

In 1978, Minott categorized female human body shapes in components to aid in patternmaking for apparel[6]. In the development her method of fitting apparel patterns, she observed shoulder and hip size considering the relationship with other body parts. Posture was also taken into account in order to adjust measurement data for more accurate patterns.

In 1981, August assessed female body shape in relation to dressmaking[7]. She developed four categories of body type designated as A, X, V, and H. These were observed from a front view of the subject. Side views were qualitatively evaluated, as well, and utilized lower case designations like b, d, i, and r to indicate categories. The August method of categorizing body shapes was based on landmark identification and recognition by component.

In 1987, Armstrong described four female body shapes based on the shoulder/hip relationship[8]. While these categories could be advantageous to patternmaking, they are limited to that application.

In 2004, Simmons, Istook, and Devarajan developed a shape sorting software[9, 10], called the Female Figure Identification Technique (FFIT) for apparel, to classify 3D body scans and identify body shapes. The measurements were 1-dimensional measures taken from 3-dimensional body scans. Using mathematical criteria and the tacit knowledge of garment design and fitting experts, a set of nine shapes was defined with mathematical descriptors.

In 2020, Sokolowski and Bettencourt modified the FFIT mathematical formulas to be more inclusive of plus size women[11].

In 2006, Connell, Ulrich, Brannon, Alexander, and Presley used experts' knowledge to develop a set of scales to assess female body shapes as visualized in body scans[12], resulting in an instrument that could be applied through software to the analysis of body scan data.

In 2009, Nakamura and Kurokawa used the 3D measurements of 560 Japanese women taken in laser metrology.[13] The data obtained for each subject consisted of approximately 160,000 body surface points. After reducing the number of data, they did a hierarchical clustering with Ward's method and obtained 5 clusters.

In 2012, Cottle[14] developed a methodology to explore body shape analysis using 3D digital data generated by the body scanner. The exploratory research design consisted of pretesting, unsupervised clustering of male body scan data, and expert recognition of male body shape clusters.

In 2013, Park and Park studied the body shapes of large people using anthropometric data from South Korea.[15] For each gender, multivariate statistical analyses were conducted to identify key factors in body shape variability and to determine representative body types.

2 Topological Data Analysis

Topological Data Analysis[16] (TDA) aims to provide mathematical, statistical and algorithmic methods for inferring, analysing and exploiting the complex topological and geometric structures underlying data often represented as point clouds in Euclidean spaces or more general metric spaces.

Many standard models are based on the following basic pipeline:

1. The input is assumed to be a finite set of points with a notion of distance - or similarity - between them.
2. A "continuous" form is constructed on the data to highlight the underlying topology or geometry.
3. Topological or geometric information is extracted from the structures built on the top of the data.
4. The extracted topological and geometric information provides new families of features and descriptors of the data. They can be used to better understand the data, especially through visualisation, or combined with other types of features for further analysis and machine learning tasks.

2.1 Metric spaces, covers, and simplicial complexes

Since topological and geometric features are generally associated with continuous spaces, data represented as finite sets of observations do not directly reveal topological information *per se*. A reasonable way to reveal the topological structure of the data is to "connect" data points that are close to each other in order to show a global continuous form underlying the data. Quantifying the notion of proximity between data points is usually done using a distance (or dissimilarity measure), and it is often convenient in TDA to consider data sets as discrete metric spaces or samples of metric spaces.

Theorem 1 (Nerve theorem) *Let $\mathcal{U} = (U_i)_{i \in I}$ be a cover of a topological space X by open sets such that the intersection of any subcollection of the U_i's is either empty or contractible. Then, X and the nerve $C(\mathcal{U})$ are homotopy equivalent.*

2.1.1 Simplicial complex

A simplicial complex is a geometric object determined by a combinatorial feature and allowing to describe certain topological spaces by generalising the notion of triangulation of a surface. Such an object is presented as a graph whose vertices are connected by edges, to which can be attached triangular faces, themselves possibly bordered by faces of higher dimension, and so on.

This structure simplifies the computation of homology groups of certain spaces such as polyhedra and certain topological varieties which admit a simplicial complex decomposition.

2.2 Homology

Homology is a classical concept in algebraic topology, providing a powerful tool for formalising and manipulating the notion of topological features of a topological space or simplicial complex in an algebraic manner. For any dimension k, the k-dimensional "holes" are represented by a vector space H_k, whose dimension is intuitively the number of these independent features.

Let K be a (finite) simplicial complex and let k be a positive integer. The space of k-chains on K, $C_k(K)$ is the set whose elements are the formal (finite) sums of the k-simplices of K. More precisely, if $\{\sigma_1, \dots, \sigma_p\}$ is the set of k-simplices of K, then any k-chain can be written as

$$c = \sum_{i=1}^{p} \varepsilon_i \sigma_i \quad \text{with} \quad \varepsilon_i \in \mathbb{Z}_2.$$

If $c' = \sum_{i=1}^{p} \varepsilon'_i \sigma_i$ is another k-chain and $\lambda \in \mathbb{Z}_2$, the sum $c + c'$ is defined as $c + c' = \sum_{i=1}^{p} (\varepsilon_i + \varepsilon'_i)\sigma_i$ and the product λc is defined as $\lambda c = \sum_{i=1}^{p} (\lambda \varepsilon_i)\sigma_i$, making $C_k(K)$ a vector space with coefficients in $\mathbb{Z}_2$. A k-chain can be seen as a finite collection of k-simplices and the sum of two k-chains as the symmetric difference of the two corresponding collections.[1]

The boundary of a k-simplex $\sigma = [v_0, \dots, v_k]$ is the $(k-1)$-chain

$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_k]$$

where $[v_0, \dots, \hat{v}_i, \dots, v_k]$ is the $(k-1)$-simplex spanned by all the vertices except v_i.[2] As the k-simplices form a basis of $C_k(K)$, ∂_k extends as a linear map from $C_k(K)$ to $C_{k-1}(K)$ called the boundary operator. The kernel $Z_k(K) = \{c \in C_k(K) \mid \partial_k = 0\}$ of ∂_k is called the space of k-cycles of K, and the image $B_k(K) = \{c \in C_k(K) \mid \exists c' \in C_{k+1}(K), \partial_{k+1}(c') = c\}$ of ∂_{k+1} is called the space of k-boundaries of K. The boundary operators satisfy

$$\partial_{k-1} \circ \partial_k \equiv 0 \quad \text{for any} \quad k \geqslant 1.$$

In other words, any k-boundary is a k-cycle, i.e. $B_k(K) \subseteq Z_k(K) \subseteq Ck(K)$.

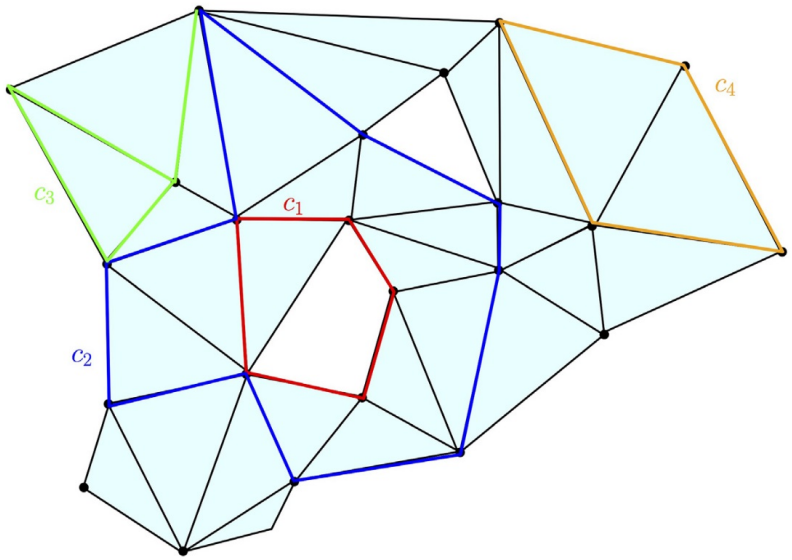


Figure 2.1

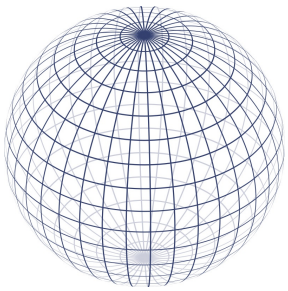
[1] The symmetric difference of two sets A and B is the set $A \Delta B = (A \setminus B) \cup (B \setminus A)$.
[2] We consider coefficients in $\mathbb{Z}_2$, here $-1 = 1$ and so $(-1)^i = 1$ for all i.

Figure 2.1: Some examples of chains, cycles, and boundaries on a two-dimensional complex K: c_1, c_2, and c_4 are one-cycles; c_3 is a one-chain but not a one-cycle; c_4 is the one-boundary, namely, the boundary of the two-chain obtained as the sum of the two triangles surrounded by c_4. The cycles c_1 and c_2 span the same element in $H_1(K)$ as their difference is the two-chain represented by the union of the triangles surrounded by the union of c_1 and c_2.

Definition 1 (Simplicial homology and Betti numbers)
The k-th (simplicial) homology group of K is the quotient vector space

$$H_k(K) = Z_k(K)/B_k(K).$$

The k-th Betti number of K is the dimension $\beta_k(K) = \dim H_k(K)$ of the vector space $H_k(K)$.



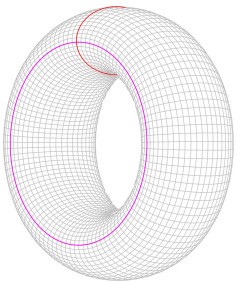
Homology

- $H_0(\mathbb{S}^2) = \mathbb{Z}$
- $H_1(\mathbb{S}^2) = \emptyset$
- $H_2(\mathbb{S}^2) = \mathbb{Z}$

Betti number

- $\beta_0(\mathbb{S}^2) = 1$
- $\beta_1(\mathbb{S}^2) = 0$
- $\beta_2(\mathbb{S}^2) = 1$

Figure 2.2: Homology groups - Sphere ($\mathbb{S}^2$)



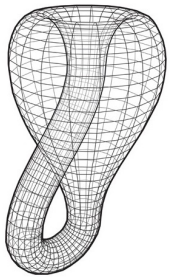
Homology

- $H_0(\mathbb{T}^2) = \mathbb{Z}$
- $H_1(\mathbb{T}^2) = \mathbb{Z} \bigoplus \mathbb{Z}$
- $H_2(\mathbb{T}^2) = \mathbb{Z}$

Betti number

- $\beta_0(\mathbb{T}^2) = 1$
- $\beta_1(\mathbb{T}^2) = 2$
- $\beta_2(\mathbb{T}^2) = 1$

Figure 2.3: Homology groups - Torus ($\mathbb{T}^2$)



Homology

- $H_0(\mathbb{K}^2) = \mathbb{Z}$
- $H_1(\mathbb{K}^2) = \mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$
- $H_2(\mathbb{K}^2) = \emptyset$

Figure 2.4: Homology groups - Klein's bottle ($\mathbb{K}^2$)

2.3 Vietoris–Rips complex

The Vietoris–Rips complex is a way of forming a topological space from distances in a set of points. It is an abstract simplicial complex that can be defined from any metric space M and distance ε by forming a simplex for every finite set of points that has diameter at most ε. In other words, it is a family of finite subsets of M, in which a subset of k points is considered to form a $(k-1)$-dimensional simplex (an edge for two points, a triangle for three points, a tetrahedron for four points, etc.); if a finite set S has the property that the distance between every pair of points in S is at most ε, then S is included as a simplex in the complex.

2.4 Čech complex

The Čech complex[17] is an abstract simplicial complex constructed from a point cloud in any metric space, which is supposed to capture topological information about the point cloud or distribution from which it is drawn. Given a finite point cloud X and an $\varepsilon > 0$, we construct the complex $\check{C}_\varepsilon(X)$ as follows: Let us take the elements of X as the set of vertices of $\check{C}_\varepsilon(X)$. Then, for every $\sigma \subset X$, let $\sigma \in \check{C}_\varepsilon(X)$ if the set of ε-balls centred on the points of σ has a non-empty intersection. In other words, the Čech complex is the nerve of the set of ε-balls centered on the points of X.

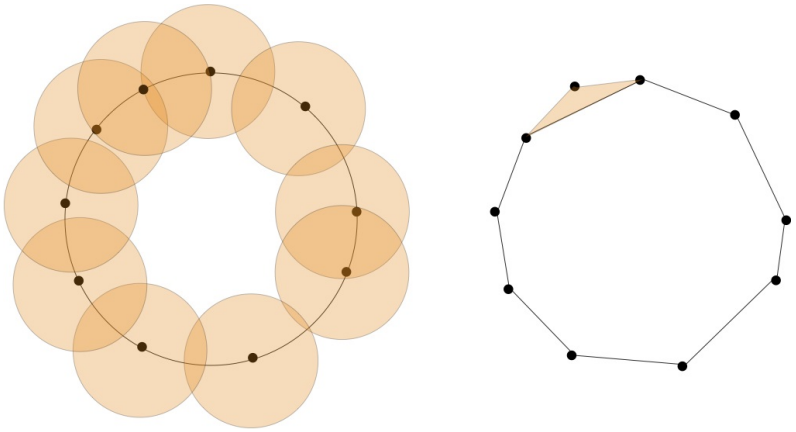


Figure 2.5: Constructing the Čech complex of a set of points sampled from a circle

The Čech complex is a subcomplex of the Vietoris-Rips complex. Although the Čech complex is more computationally expensive than the Vietoris-Rips complex, since we have to check the higher-order intersections of the balls of the complex, the Nerve Theorem guarantees that the Čech complex is homotopically equivalent to the union of the balls of the complex while the Vietoris-Rips complex may not be.

2.5 Persistence module

A persistence module $\mathbb{U}$ indexed by $\mathbb{Z}$ is a vector space U_t for each $t \in \mathbb{Z}$, and a linear map $u_t^s : U_s \to U_t$ whenever $s \leqslant t$, such that $u_t^t = 1$ for all t and $u_t^s u_s^r = u_t^r$ whenever $r \leqslant s \leqslant t$.[18] An equivalent definition is a functor from $\mathbb{Z}$ considered as a partially ordered set to the category of vector spaces.

2.6 Persistent homology

Persistent homology is a tool used to efficiently compute, study and encode the multiscale topological features of nested families of simplicial complexes and topological spaces.

2.6.1 Filtrations

A filtration of a simplicial complex K is a nested family of subcomplexes $(K_r)_r \in T$, where $T \subseteq \mathbb{R}$, such that for any $r, r' \in T$, if $r \leqslant r'$ then $K_r \subseteq K_{r'}$ and $K = \cup_{r \in T} K_r$. The subset T can be either finite or infinite.

In practice, the parameter $r \in T$ can often be interpreted as a scaling parameter, and the filtrations classically used in TDA often belong to one of the following two families.

Filtrations built on bop of data

Given a subset $\mathbb{X}$ of a compact metric space (M, ρ), the families of Vietoris–Rips complexes $(Rips_r(\mathbb{X}))_{r \in \mathbb{R}}$ and Čech complexes $(Cech_r(\mathbb{X}))_{r \in \mathbb{R}}$ are filtrations.[3] Here, the parameter r can be interpreted as a resolution at which the data set $\mathbb{X}$ is considered.

Sublevel sets filtrations

Let K be a simplicial complex with vertex set V and $f : V \to R$. Then f can be extended to all simplexes of K by $f([v_0, \ldots, v_k]) = \max\{f(v_i) : i = 1, \ldots, k\}$ for any simplex $\sigma = [v_0, \ldots, v_k] \in K$ and the family of subcomplexes, $K_r = \{\sigma \in K \mid f(\sigma) \leqslant r\}$, defines a filtration called the filtration of sublevel sets of f. Similarly, we can define the filtration of the top-level sets of f.

2.6.2 Persistent modules and Persistence diagrams

Let $Filt = (F_r)_{r \in T}$ be a filtration of a simplicial complex or topological space. Given a non-negative integer k and considering the homology groups $H_k(F_r)$, we obtain a sequence of vector spaces where the inclusions $F_r \subset F_{r'}$, $r \leqslant r'$ induce linear maps between $H_k(F_r)$ and $H_k(F_{r'})$. Such a sequence of vector spaces together with the linear maps connecting them is called a persistence module.

Definition 2 *A persistence module $\mathbb{V}$ over a subset T of $\mathbb{R}$ is an indexed family of vector spaces $(V_r | r \in T)$ and a doubly indexed family of linear maps $(v_s^r : V_r \to V_s | r \leqslant s)$ which satisfy the composition law $v_t^s \circ v_s^r = v_t^r$ whenever $r \leqslant s \leqslant t$, and where v_r^r is the identity map on V_r.*

When a persistence module $\mathbb{V}$ can be decomposed as a direct sum of interval modules, one can show[19] that this decomposition is unique up to reordering the intervals. As a consequence, the set of resulting intervals is independent of the decomposition of $\mathbb{V}$ and is called the persistence barcode of $\mathbb{V}$. The disjoint union of these points, together with the diagonal $\Delta = \{x = y\}$, is a multi-set called the persistence diagram of $\mathbb{V}$.

[3] By convention, for $r < 0$, $Rips_r(\mathbb{X}) = Cech_r(\mathbb{X}) = \emptyset$.

Example

In figure 2.6: (**A**) For radius $r = 0$, the union of balls is reduced to the initial finite set of points, each of which corresponds to a zero-dimensional feature, i.e. a connected component; an interval is created for the birth of each such feature at $r = 0$. (**B**) Some of the balls began to overlap, causing some of the related components to die and merge; the persistence diagram keeps track of these deaths, ending the corresponding intervals as they disappear. (**C**) New components merged, giving rise to a single related component and thus all intervals associated with a unidimensional feature ended, except for the one corresponding to the remaining components; two new unidimensional features emerged, giving rise to two new intervals (in blue) starting on the birth scale. (**D**) One of the two one-dimensional cycles has been filled, resulting in its death in the filtration and the end of the corresponding blue interval. (**E**) All one-dimensional features are dead; only the long (and never dying) red interval remains. The final barcode can also be represented equivalently as a persistence diagram where each interval (a, b) is represented by the coordinate point $(a, b) \in \mathbb{R}^2$.

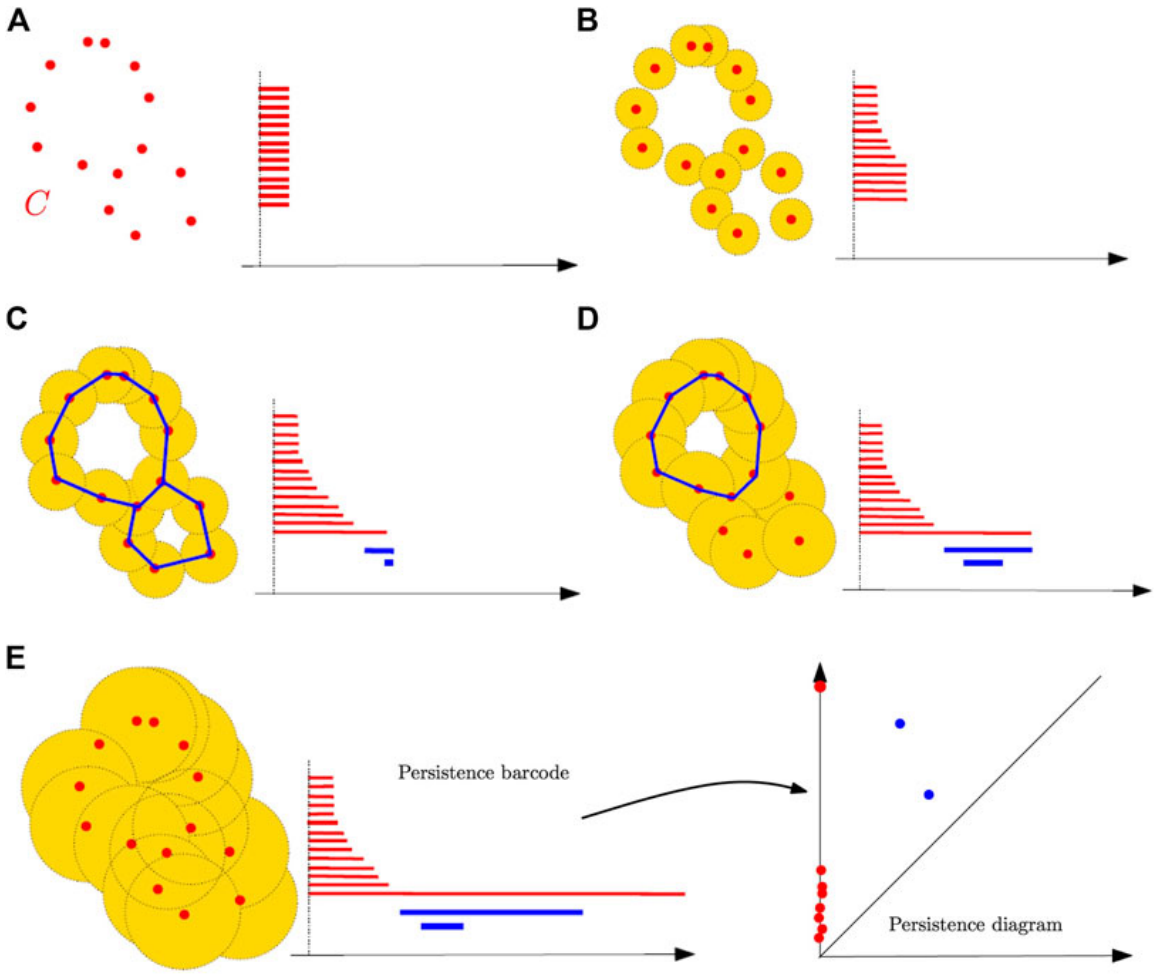


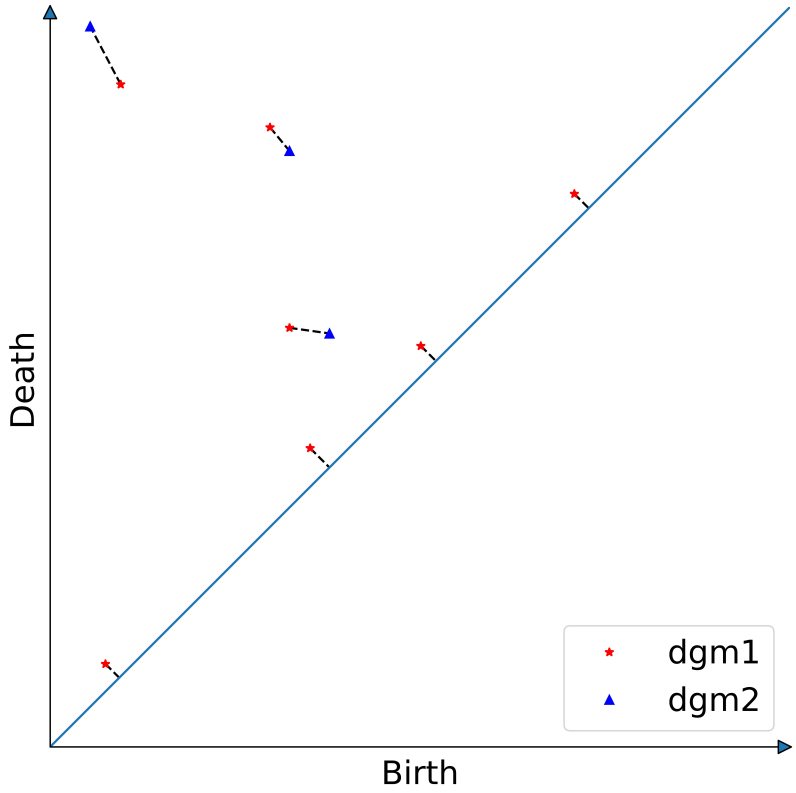
Figure 2.6: The filtering by set of sublevels of the distance function to a scatterplot and the construction of its persistence barcode as the radius of the balls increases. The blue curves represent the monocycles associated with the blue bars in the barcodes. The persistence diagram is finally defined from the persistence barcodes.

Intuitively, the longer an interval in the barcode or, equivalently, the further the corresponding point in the diagram is from the diagonal, the more persistent, and therefore relevant, the corresponding homological feature is through the filtration.

It is also worth noting that for a given radius r, the k-th Betti number of the corresponding ball union is equal to the number of persistence intervals corresponding to k-dimensional homological features and containing r. Thus, the persistence diagram can be seen as a multiscale topological signature encoding the homology of the ball union for all radii as well as its evolution through the values of r.

2.7 Wasserstein Distance

The q-Wasserstein distance measures the similarity between two persistence diagrams.



The **Wasserstein distance** between two persistence diagrams X and Y is defined as

$$W_p^q(X,Y) := \inf_{\varphi:X\to Y} \left(\sum_{x\in X} (\|x-\varphi(x)\|_q)^p \right)^{\frac{1}{p}}$$

where $1 \leqslant p, q \leqslant \infty$ and φ ranges over bijections between X and Y.
The **bottlneck distance** between X and Y is

$$W_\infty^q(X,Y) := \inf_{\varphi:X\to Y} \sup_{x\in X} \|x-\varphi(x)\|_q$$

This is a special case of Wasserstein distance, letting $p = \infty$.

3 The CAESAR dataset (UCSC)

In 2014, Yang, Yu, Zhou, Du, Davis and Yang[20] developped a novel approach to generate human body models in a variety of shapes and poses via tuning semantic parameters. Their approach was investigated with datasets of up to 3000 scanned body models (CAESAR) which have been placed in point to point correspondence. Correspondence is established by nonrigid deformation of a template mesh.

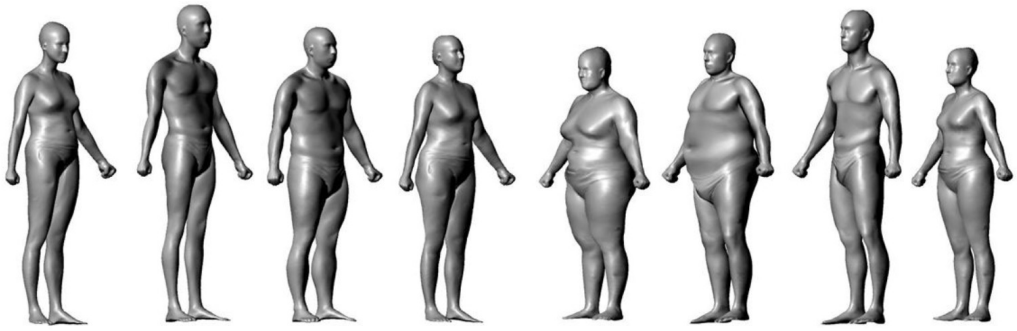


Figure 3.1: By fitting a template to many raw range scans, meshes with a wide variety of body shapes having point-to-point correspondences can be obtained.

The data is derived from the CAESAR dataset.
We use this dataset which contains about 1500 registered male and female meshes with point-to-point correspondences respectively.
Each mesh has 12500 vertices and 25000 facets.

3.1 Preprocessing

For each mesh, we extract the 12500 points, then we normalize its size to 170 centimeters and we compute its persistence diagram with the Python library Gudhi with a minimum persistence of 3.

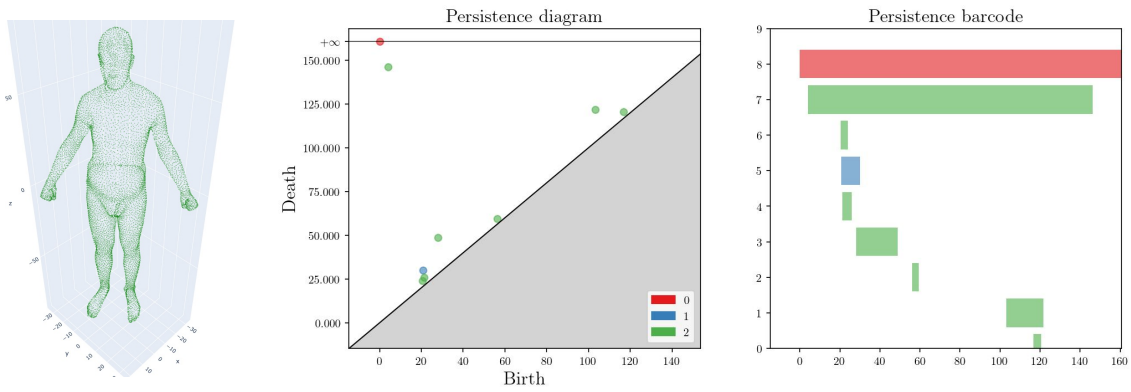


Figure 3.2: Example of a point set with its persistence diagram and persistence barcode.

Then we calculate a distance matrix where the distance between two digrams is calculated as the square root of the sum of the squares of the Wasserstein distances of each homology group (H_0, H_1 and H_2) with $p = q = 2$. This matrix will be used for clustering. As the notion of barycentre is not perfectly defined, we will use clustering methods that are not based on it, i.e. hierarchical clustering and K-Medoids.

3.2 Visualisation of homologies

As the search for minimal cycles in an undirected graph is an NP-complete problem, our first idea was to do a sub-sampling to reduce the number of points and thus decrease the computation time.

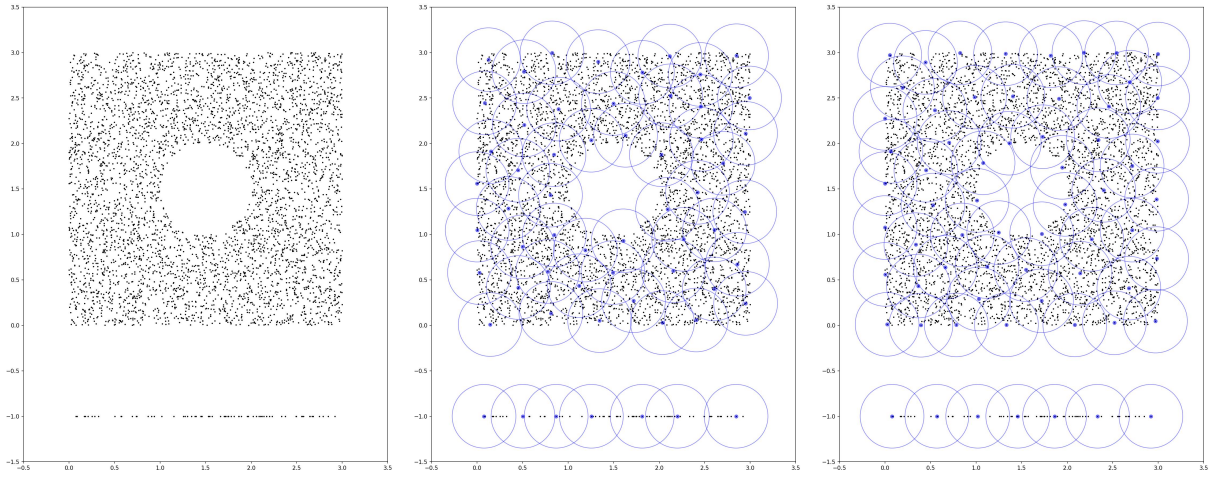
Pour cela nous nous sommes basés sur la méthode ASAP[21] (A Sub-sampling Approach for Preserving topological structures). It is an efficient sub-sampling strategy inspired by Coulomb's law to decrease the number of data points in d-dimensional point clouds while preserving its homology.

3.2.1 ASAP

Here is the algorithm which, from a point set N, calculates a sub-sample N_r.
This consists of randomly selecting points separated by a distance greater than a radius r and then using Coulomb's law several times, considering all the selected points as elements of the same charge, in order to homogenise the subsample.

Data: N, radius r, $N_r = \emptyset$, $\gamma = 1$ et $t = 1$
Result: N_r
while $N_{tmp} \neq \emptyset$ **do**
  Pick a point s_i at random in N_{tmp}
  $N_r \leftarrow N_r \cup \{s_i\}$ and $N \leftarrow N \backslash \mathcal{B}(s_i, r)$
end
while $\gamma > 0.1r^3$ **do**
  $\gamma = r^3 \exp(-t/\tau)$
  foreach $s_i \in N_r$ **do**
    $m_i = \sum_{s_j \in \mathcal{N}_i} \frac{\gamma(s_j - s_i)}{\|s_j - s_i\|^3}$ and $\hat{s}_i = \arg\min_{p_j}(d(p_j, s_i + m_i)) \quad \forall p_j \in N$
    if $d(s_i, \hat{s}_i) > r \forall s_j \in N_r$ **and** $s_j \neq s_i$ **then**
      $s_i = \hat{s}_i$
    end
  end
  $N_{tmp} = N$
  foreach $s_i \in N_r$ **do**
    Remove all points belonging to $\mathcal{B}(s_i, r)$ from N_{tmp}.
  end
  while $N_{tmp} \neq \emptyset$ **do**
    Pick a point s_i at random in N_{tmp}
    $N_r \leftarrow N_r \cup \{s_i\}$ and $N \leftarrow N \backslash \mathcal{B}(s_i, r)$
  end
  $t++$
end

Algorithm 1: ASAP algorithm

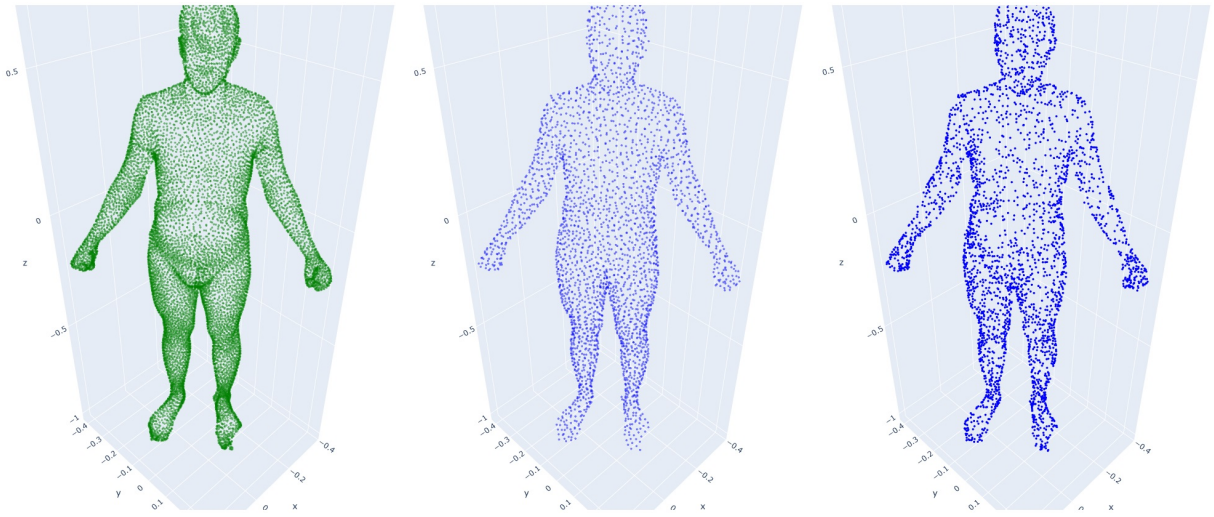


(a) (b) (c)

Figure 3.3: (a) Points N distributed on a line and holed square, (b) ball cover after random sub-sampling and (c) after repulsive selection.

Example on a body point set

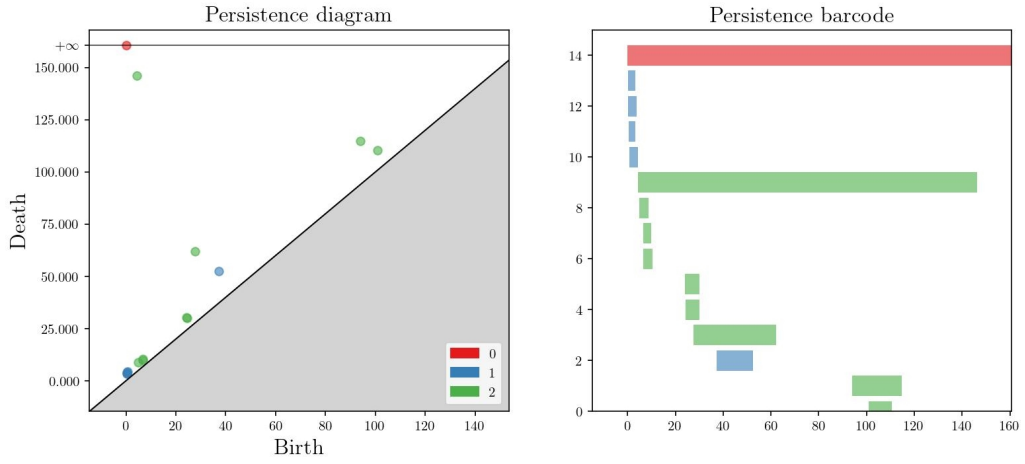
Here is an example of the ASAP subsample on sample number 1487 of the male mesh set. The radius used is the square root of the minimum persistence used during pre-processing, i.e. $\sqrt{3}$.



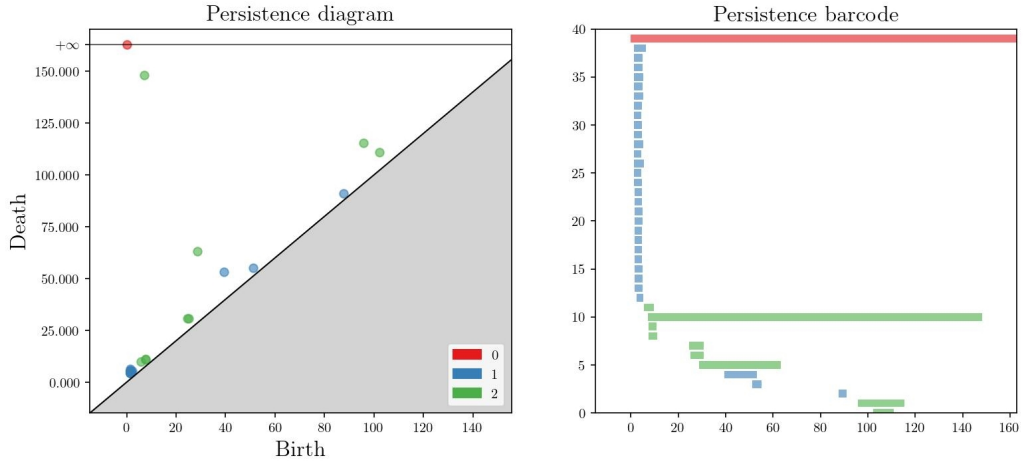
(a) (b) (c)

Figure 3.4: (a) Original point set (b) ASAP subsample and (c) random subsample.

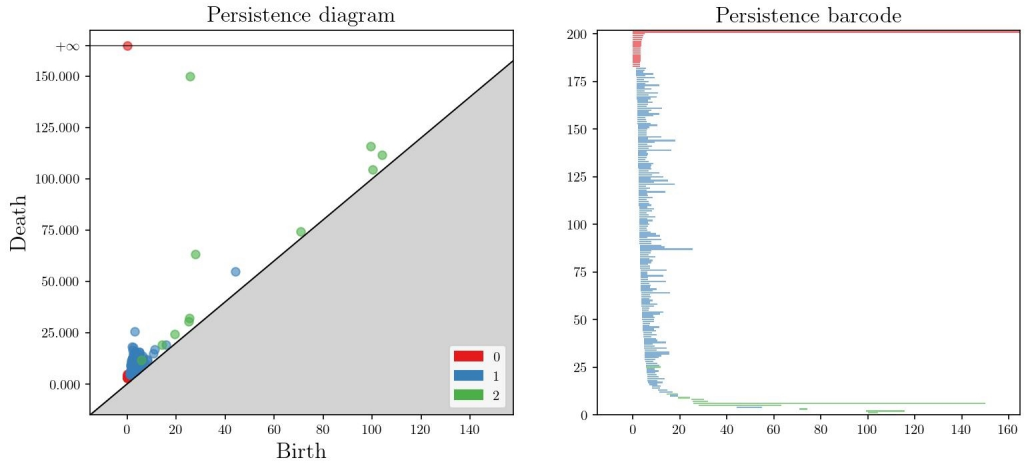
The original set consists of 12500 points and the ASAP subsample consists of only 2957 points. As we can see from the figure above, the ASAP subsample looks very similar to the original sample and is much more homogeneous than a random subsample.



(a)



(b)



(c)

Figure 3.5: (a) Original point set (b) ASAP subsample and (c) random subsample.

Here, the number of H_1 homologies increases reasonably with ASAP sub-sampling while H_0 and H_2 homologies are rather well preserved. Whereas for the random subsample the number of H_0 and H_1 homologies increases strongly and H_2 homologies seem less well preserved.
This can also be seen with the distances to the original of the subsamples:

- $W_2^2(Original, ASAP) \simeq 13.69$
- $W_2^2(Original, Random) \simeq 64.48$

Unfortunately, the ASAP sub-sampling does not allow to reduce the computational cost of the minimum cycle search sufficiently, except by reducing the number of points drastically, but this does not allow to keep the topological information of the original set. We therefore need another approach.

3.2.2 Make it simple

Since displaying the homologies in their entirety is too costly, we thought of other approaches for each degree of homologies.

For each homology, we know the radii of the balls at its birth and death.

Simplex tree

A simplex tree[22] represents abstract simplicial complexes of any dimension. All faces of the simplicial complex are explicitly stored in a trie whose nodes are in bijection with the faces of the complex. This data structure allows to efficiently implement a large range of basic operations on simplicial complexes.

Using the simplex tree of a set of points, we know the values of the radii when pairs of points, triangles and tetrahedra are covered.

- **H_0 homologies** All H_0 homologies are born when the radius of the balls is zero. For each homology H_0, we choose to display the second point of the pair covered at the birth of the homology as its representative.

- **H_1 homologies** First, we make an undirected graph containing all the points of a set, where each time a pair of points is covered, as the radius of the balls increases, we connect these points by an edge with a weight equal to the radius of the balls. At the birth of a homology H_1, before adding the edge to our graph, we compute the shortest path connecting these two points which we display by closing it with the segment connecting these points.
  The lace displayed is a likely representative of this homology.
  At the death of this homology, we recover the information of the triangle covered by the balls and we add it to the display to give a general idea of the evolution of our homology.

- **H_2 homologies** For each homology H_2, we simply display the triangle covered at its birth and the tetrahedron covered at its death.

Example of homology display

Here are two examples of homology posters for sample number 1487: The fifth homology H_1 (fig. 3.6) corresponding to the inter-ankle surface and the second homology H_2 (fig. 3.7) corresponding to the right hand volume of the sample.

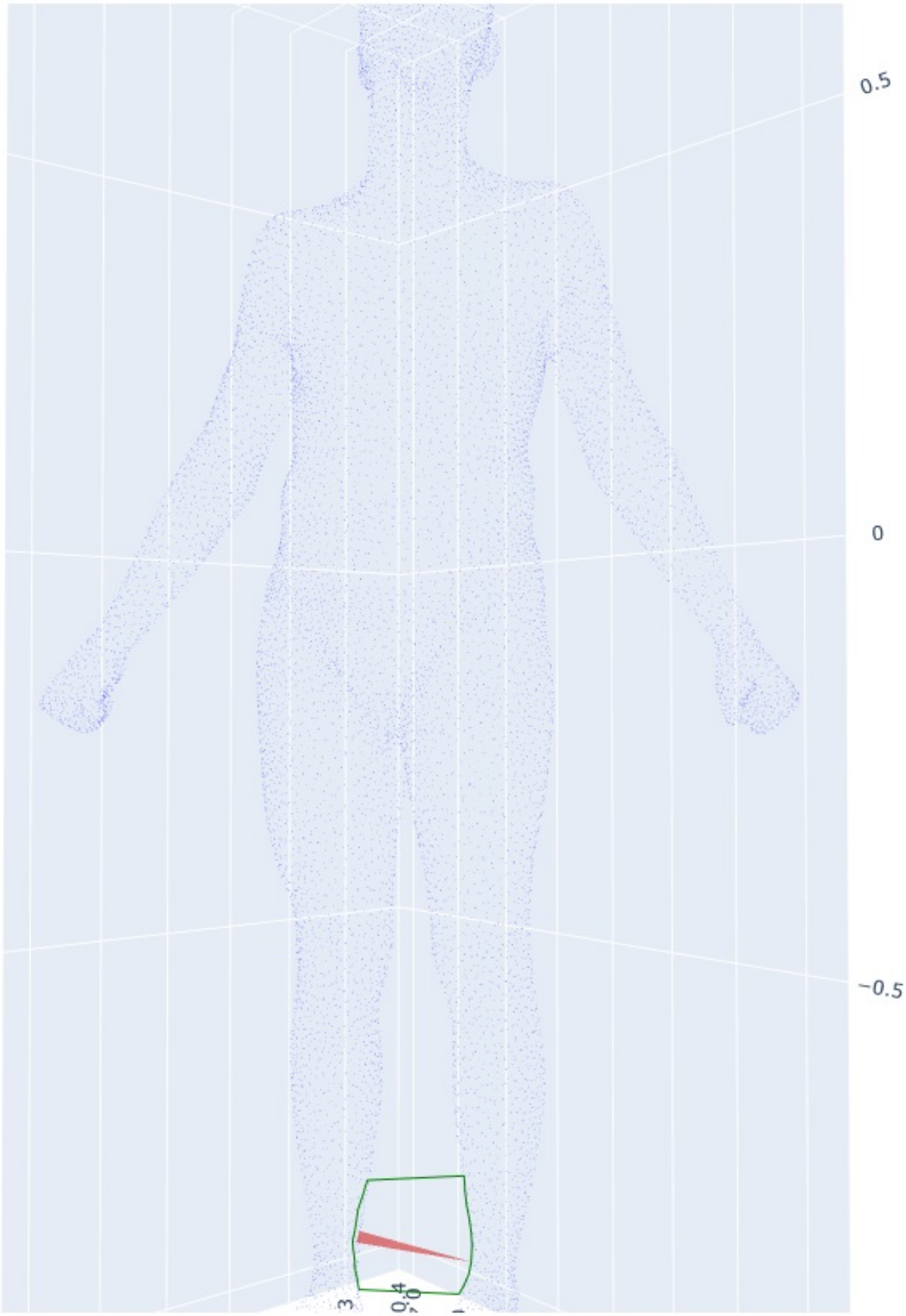


Figure 3.6: H1 n° 5

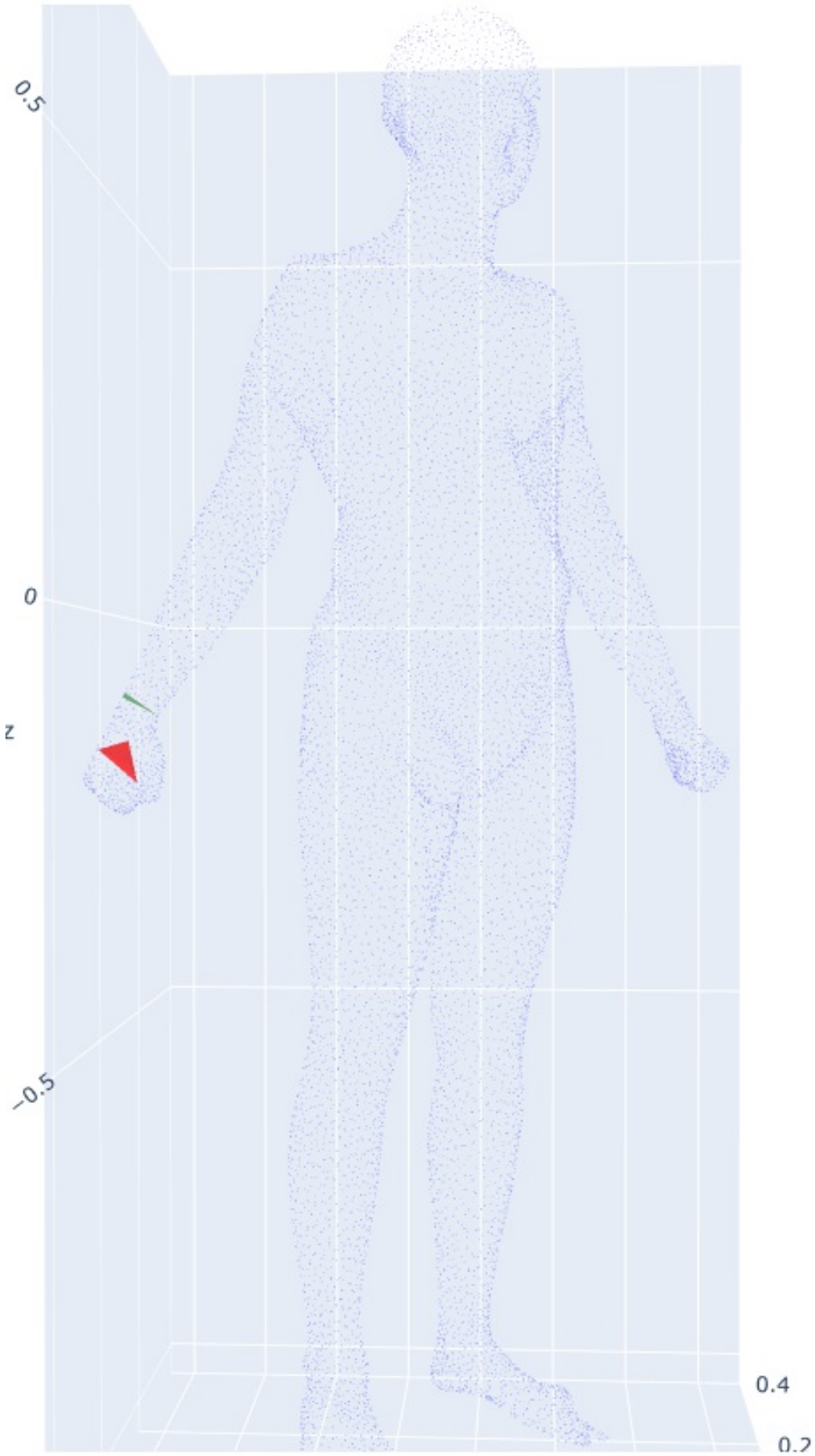


Figure 3.7: H2 n° 2

3.3 Clustering

We will first present our results on the male meshes, then on the female meshes.

3.3.1 Male dataset

First, we compute the pair of farthest and closest diagrams, to visually check the Wasserstein distance.

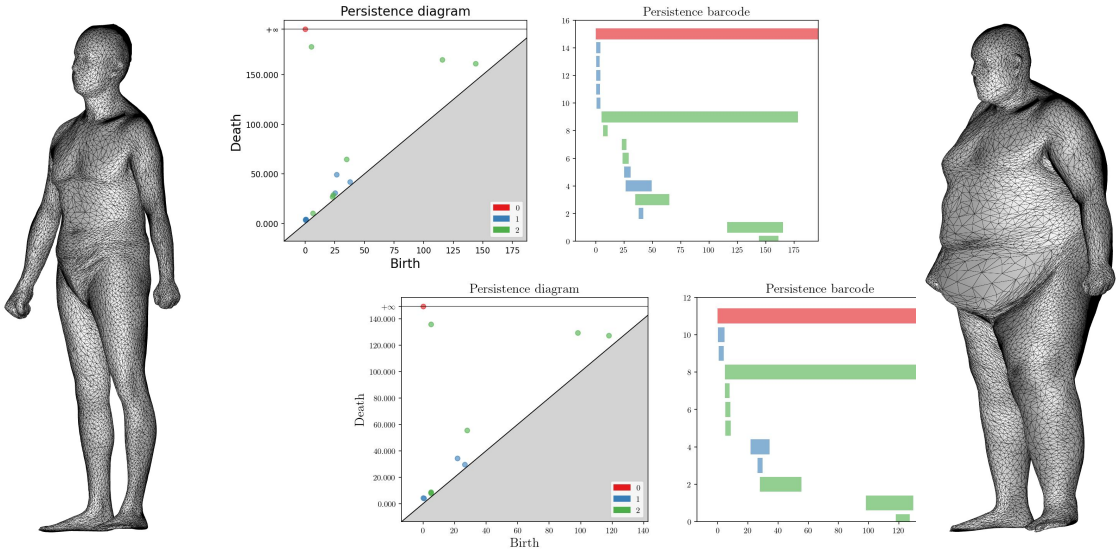


Figure 3.8: Farthest diagrams and their scans

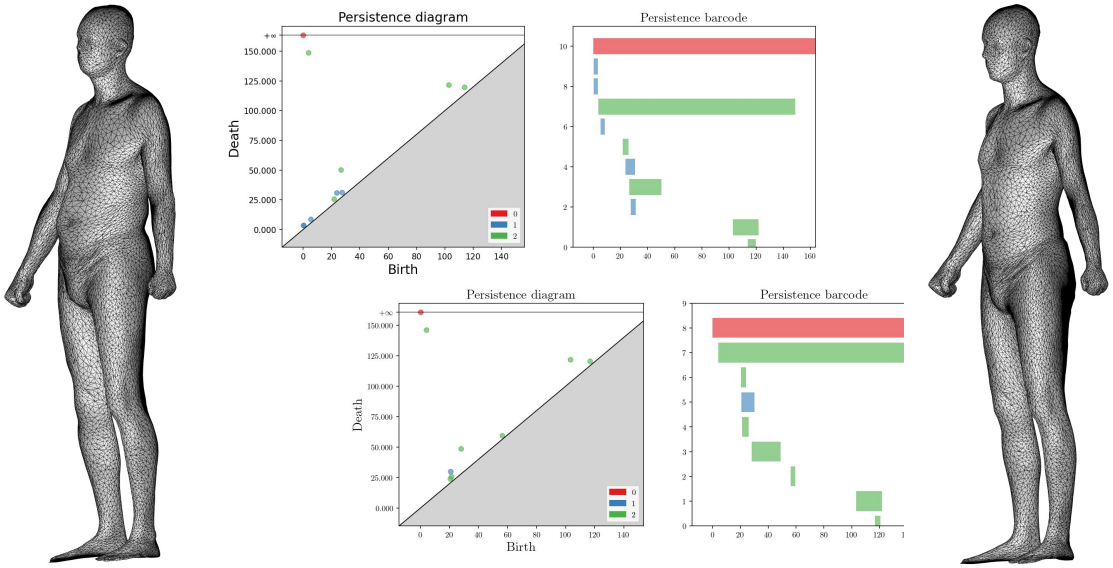


Figure 3.9: Closest diagrams and their scans

As we do not have the notion of a barycentre, we use the 'complete' linkage method.

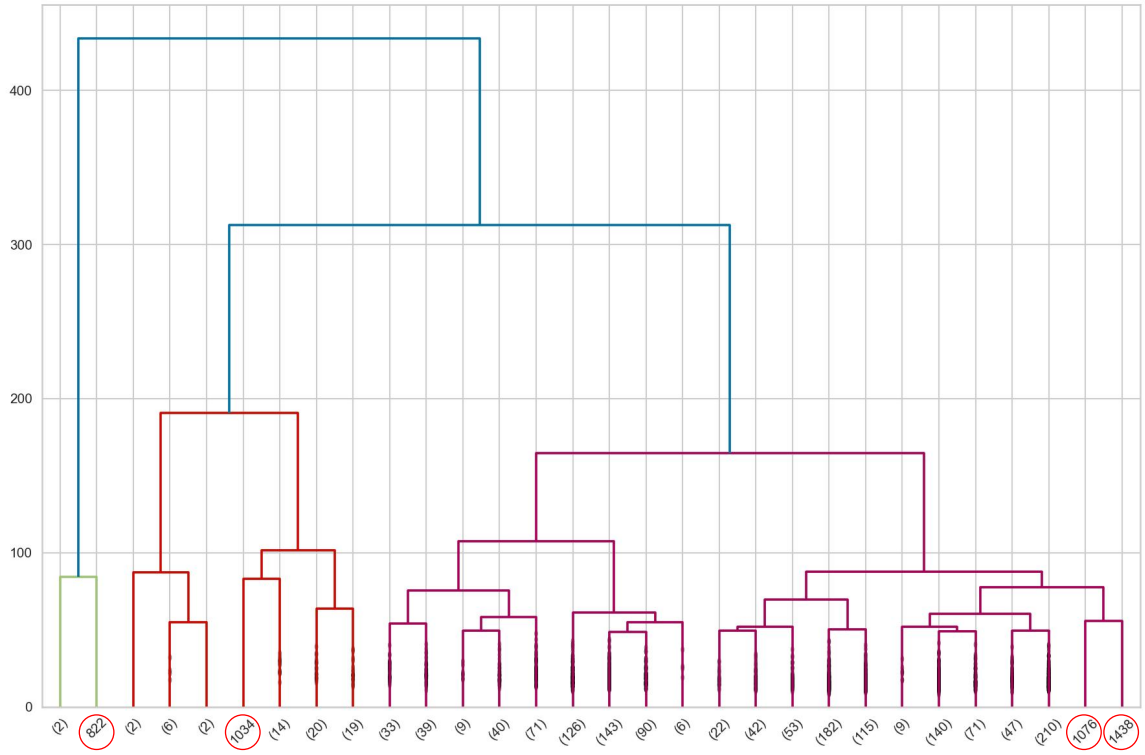


Figure 3.10: Dendrogram with 'complete' linkage

Diagrams that are on their own in a sheet of the dendrogram are all associated with faulty scans.

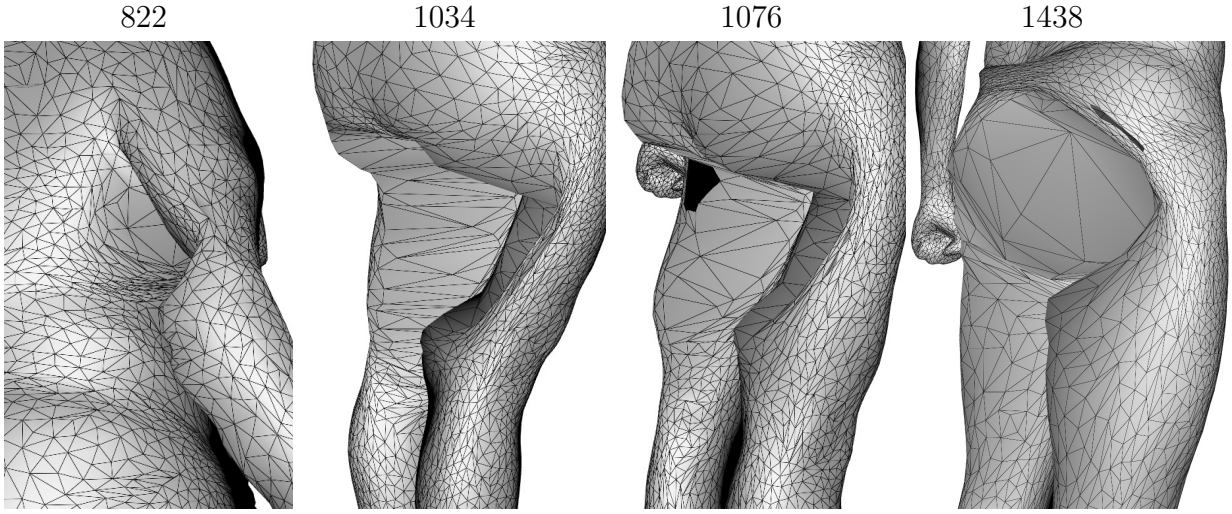


Figure 3.11: Faulty scans detected with 'complete' linkage

We use the elbow curve to determine the optimal number of clusters.

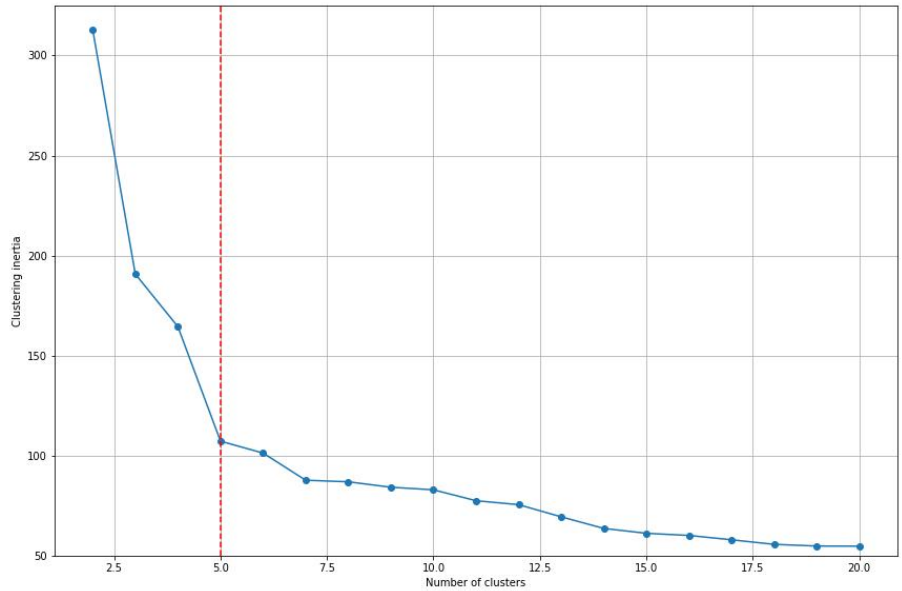


Figure 3.12: Elbow curve for complete linkage hierarchical clustering.

From the figure above, the optimal number of clusters seems to be 5.

Cluster n°	1	2	3	4	5
Size	3	10	54	557	893
Proportion	0.2%	0.66%	3.56%	36.72%	58.87%

Table 3.1: Clustering results

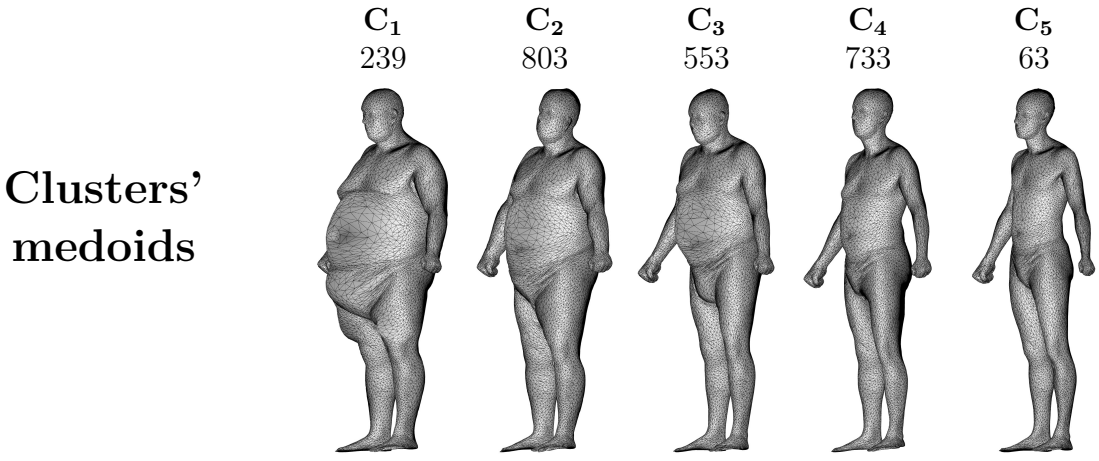


Figure 3.13: Indexes and scans of medoids.

Let C_k be cluster number k.
According to the cluster representatives, C_1 contains the 3 largest individuals and as k increases, C_k contains thinner and more numerous individuals.

3.3.2 Female dataset

First, we compute the pair of farthest and closest diagrams.

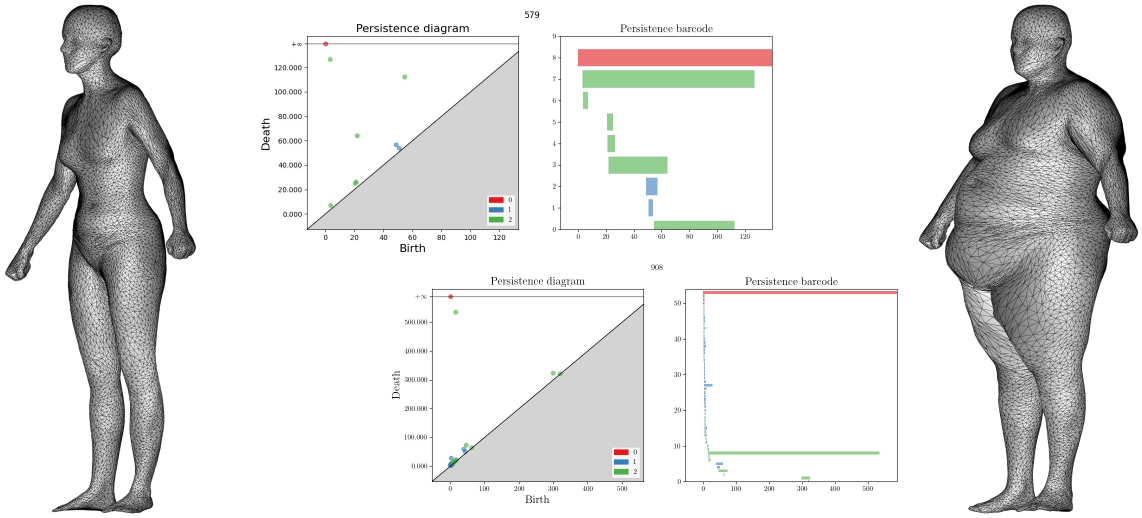


Figure 3.14: Farthest diagrams and their scans

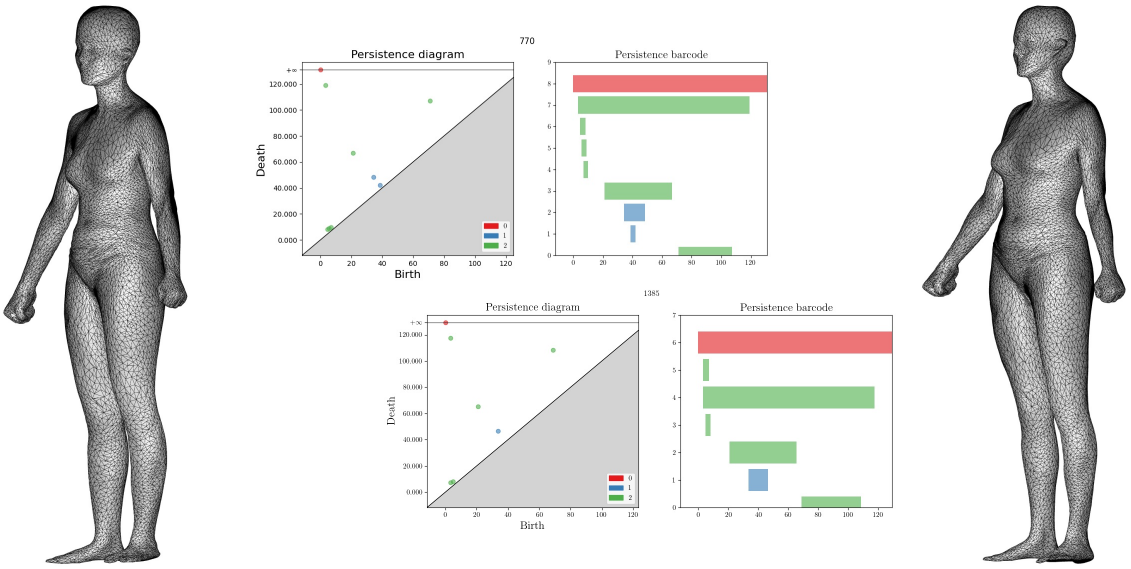


Figure 3.15: Closest diagrams and their scans

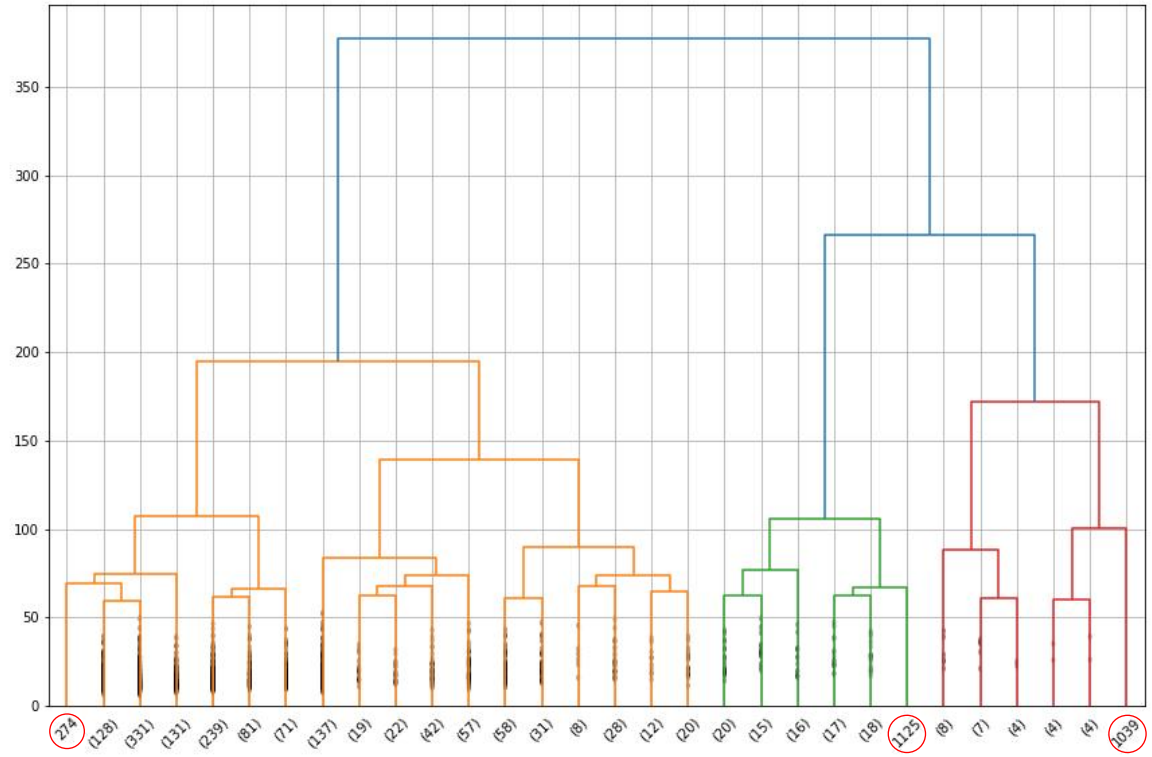


Figure 3.16: Dendrogram with 'complete' linkage

As for males, diagrams that are on their own in a sheet of the dendrogram are all associated with faulty scans.

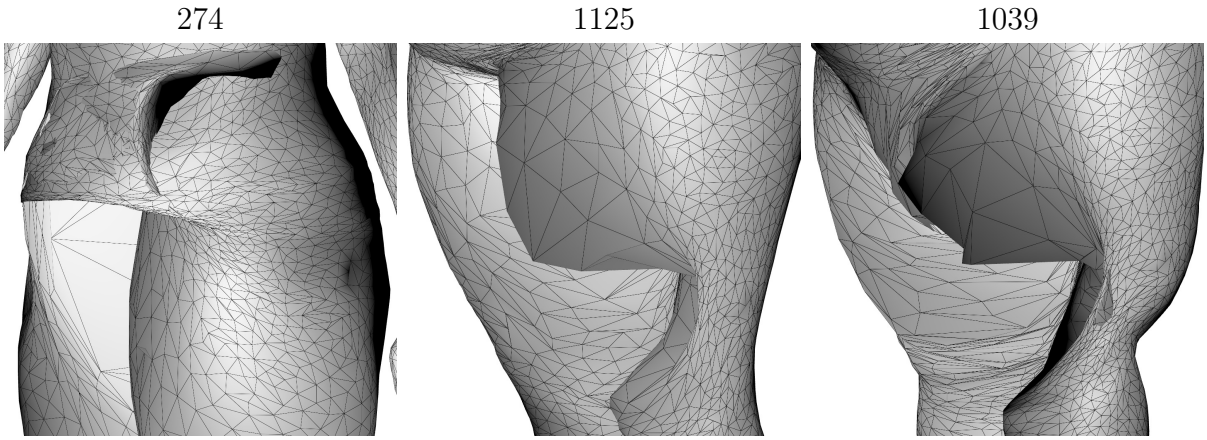


Figure 3.17: Faulty scans detected with 'complete' linkage

We use the elbow curve to determine the optimal number of clusters.

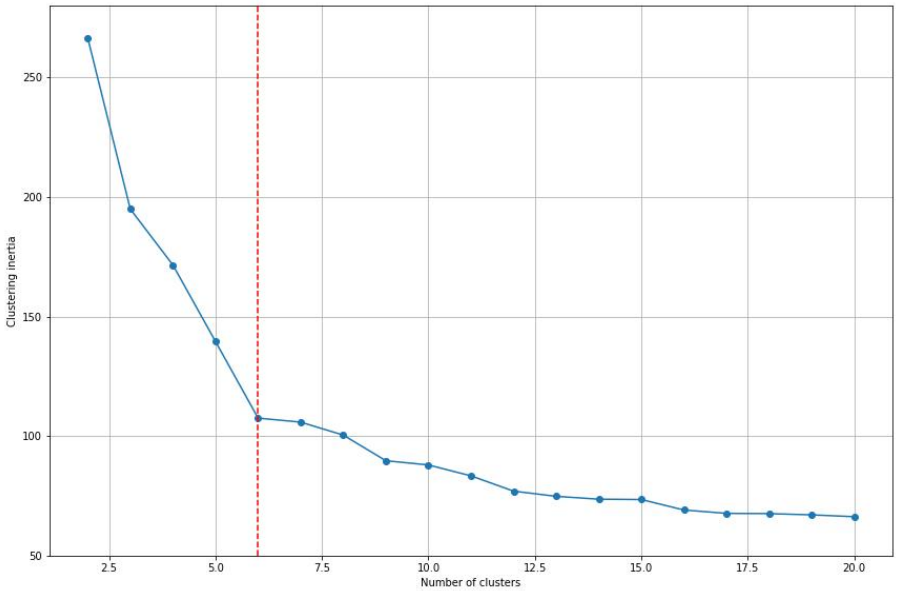


Figure 3.18: Elbow curve for complete linkage hierarchical clustering.

From the figure above, the optimal number of clusters seems to be 6.

Cluster n°	1	2	3	4	5	6
Size	982	277	157	87	19	9
Proportion	64.14%	18.09%	10.25%	5.68%	1.24%	0.59%

Table 3.2: Clustering results

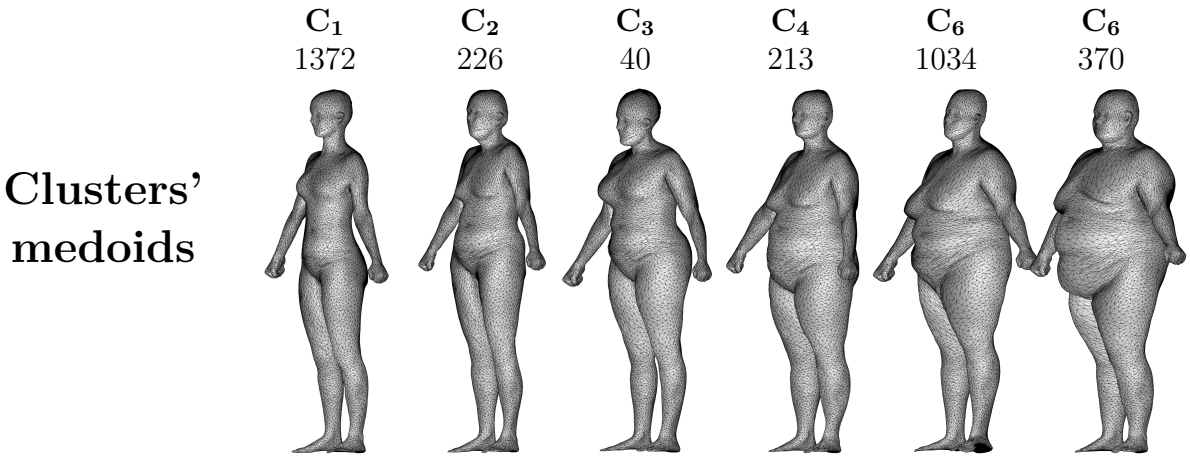


Figure 3.19: Indexes and scans of medoids.

Let C_k be cluster number k.
According to the cluster representatives, C_6 contains the 9 largest individuals and as k decreases, C_k contains thinner and more numerous individuals.

3.3.3 Hierarchical clustering with 'Ward' linkage

Although we do not have a notion of barycentre of persistence digrams, we can still use the Ward linkage thanks to the Lance-Williams[23] algorithm.

Ward linkage applied to female diagrams

Let us test the hierarchical clustering with Ward linkage on the set of female diagrams. We choose to make 6 clusters as with the complete linkage.

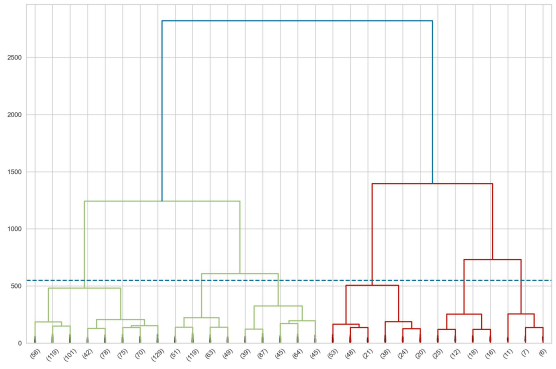


Figure 3.20: Dendrogram

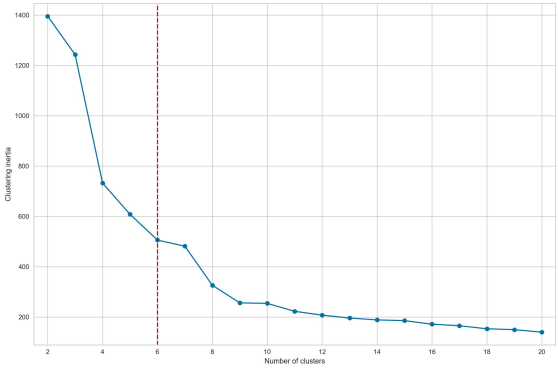


Figure 3.21: Elbow curve

We use the dendrogram and the elbow curve to determine the optimal number of clusters. The optimal number of clusters seems to be 6.

Cluster n°	1	2	3	4	5	6
Size	670	282	280	204	71	24
Proportion	43.76%	18.42%	18.29%	13.32%	4.64%	1.57%

Table 3.3: Clustering results

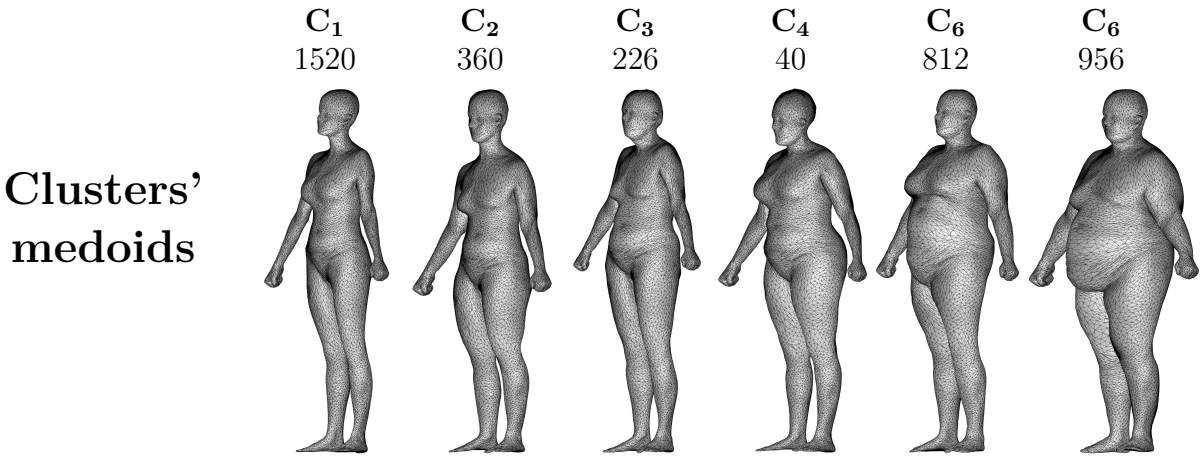


Figure 3.22: Indexes and scans of medoids.

Let C_k be cluster number k.
We get similar results to those obtained with the complete linkage, C_6 contains the 24 largest individuals and as k decreases, C_k contains thinner and more numerous individuals. Cluster sizes are less disparate than for complete linking.

3.4 Persistence landscapes

Persistence landscapes are an encoding of persistence diagrams by a series of piecewise continuous linear functions. This allows us to perform statistics on them, the absence of which was a disadvantage of persistence diagrams. In particular, it is possible to calculate (unique) averages of landscapes. While a persistence landscape has a corresponding persistence diagram, an average of persistence landscapes does not.

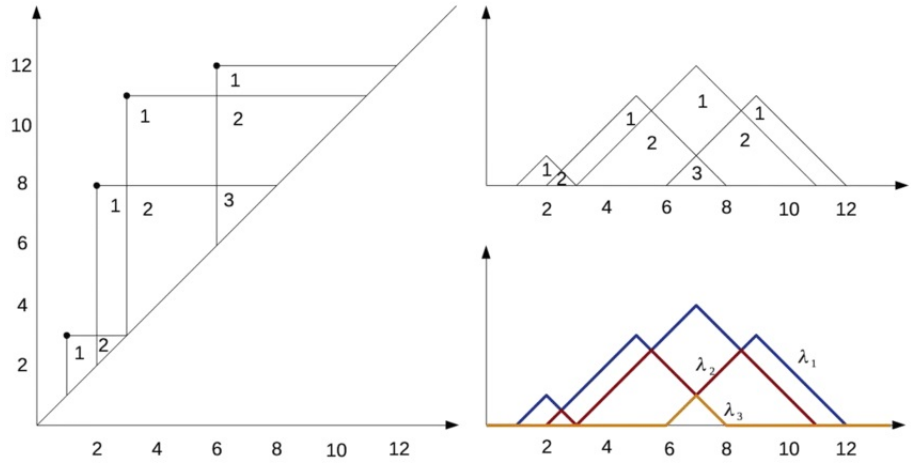
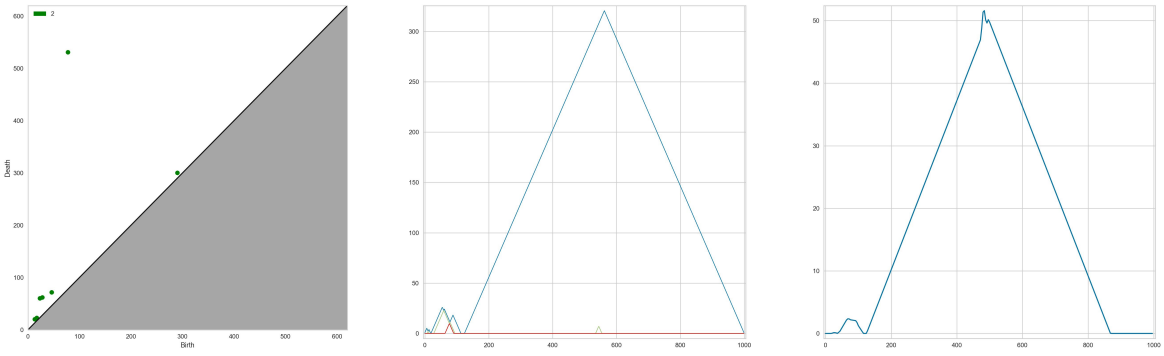


Figure 3.23: A visual explanation of persistence landscapes[24]. The persistence diagram (left) is tilted, so that the diagonal becomes the new horizontal axis (top right). The λ_i are the piecewise linear functions (bottom right)

3.4.1 Persistence silhouette

A persistence silhouette is computed by taking a weighted average of the collection of 1D piecewise-linear functions given by the persistence landscapes, and then by evenly sampling this average on a given range. Finally, the corresponding vector of samples is returned.

Figure 3.24: Example



Persistence diagram Persistence landscape Persistence silhouette

Clustering

For the implementation of clustering, we choose to make a vector consisting of 25 points of the silhouette of H_0 homologies, 250 points equidistant from the silhouette of H_1

homologies and 250 points equidistant from the silhouette of H_2 homologies, for each persistence diagram. The points are the values of the silhouette equally spaced.

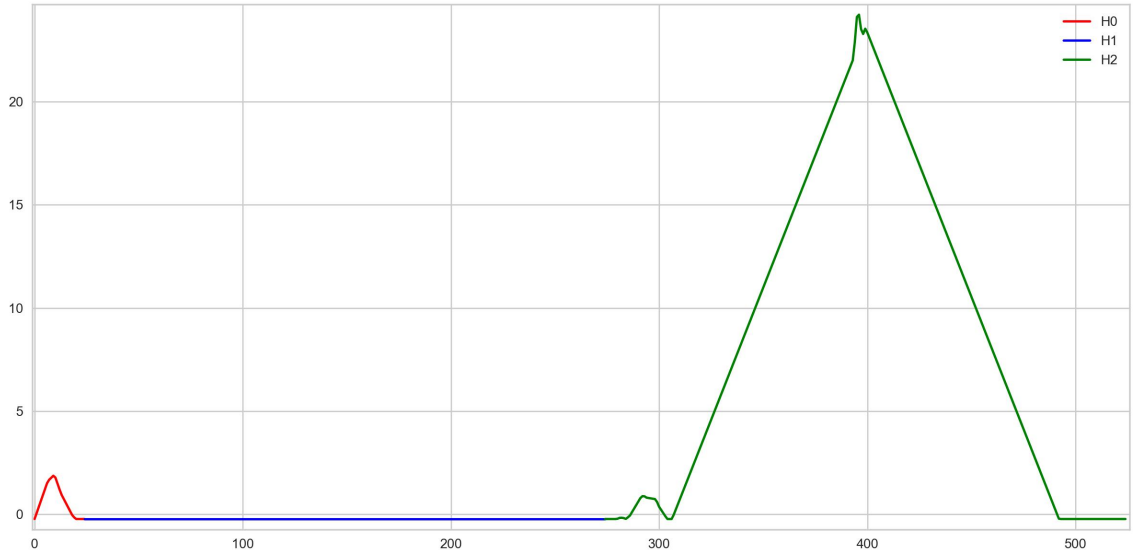


Figure 3.25: Representation of a vector

3.5 Gender discrimination as a quality index for clustering.

Although some morphotype groups may contain both males and females (e.g. obese or very thin people), it is expected that the majority of clusters obtained by mixing female and male diagrams will consist overwhelmingly of one or the other gender.

Let $P_m(C)$ and $P_f(C)$ be the proportions of male and female diagrams in a cluster C, respectively.

$$P_m(C) = \frac{n_m(C)}{s(C)}$$
$$P_f(C) = \frac{n_f(C)}{s(C)}$$
$$P_m(C) + P_f(C) = 1$$

where $n_m(C)$ is the number of male diagrams in C, $n_f(C)$ the number of female diagrams in C and $s(C)$ is the size of C.

To check the quality of the $\mathcal{C}$ clustering of a set of mixed D_{MF} diagrams, we introduce a gender discrimination index (GDI) such that

$$\text{GDI}(\mathcal{C}) = \frac{2}{s(D_{\text{MF}})} \sum_{k=1}^{K} s(C_k) \left| P_m(C_k) - \frac{1}{2} \right|$$

where K is the number of clusters of $\mathcal{C}$, C_k are the clusters of $\mathcal{C}$ and $s(D_{\text{MF}})$ is the number of diagrams in D_{MF}.

Thus, the better the clustering $\mathcal{C}$, the closer $\text{GDI}(\mathcal{C})$ is to 1; and the worse it is, the closer $\text{GDI}(\mathcal{C})$ is to 0. We can consider that a clustering is satisfactory if its GDI is greater than or equal to $\frac{1}{2}$.

3.5.1 Wasserstein distance

Hierarchical clustering with 'complete' linkage

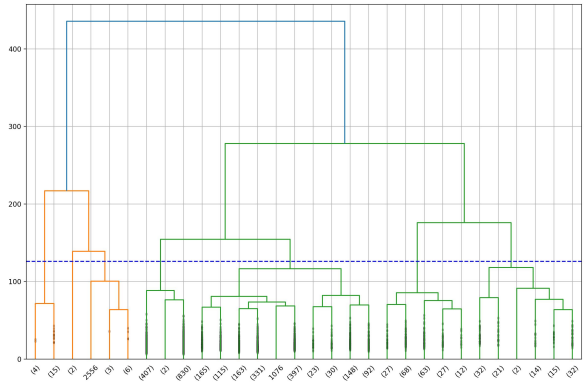


Figure 3.26: Dendrogram

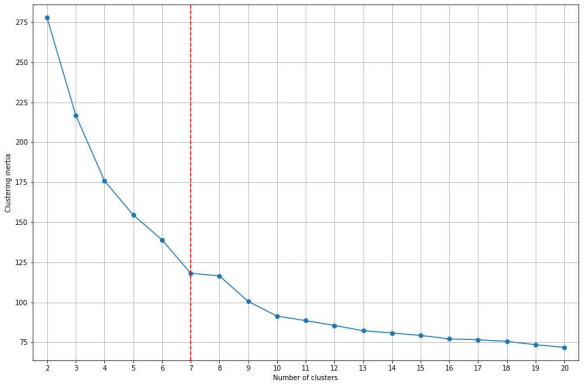


Figure 3.27: Elbow curve

According to the dendrogram and the elbow curve, the appropriate number of clusters to choose would be 5.

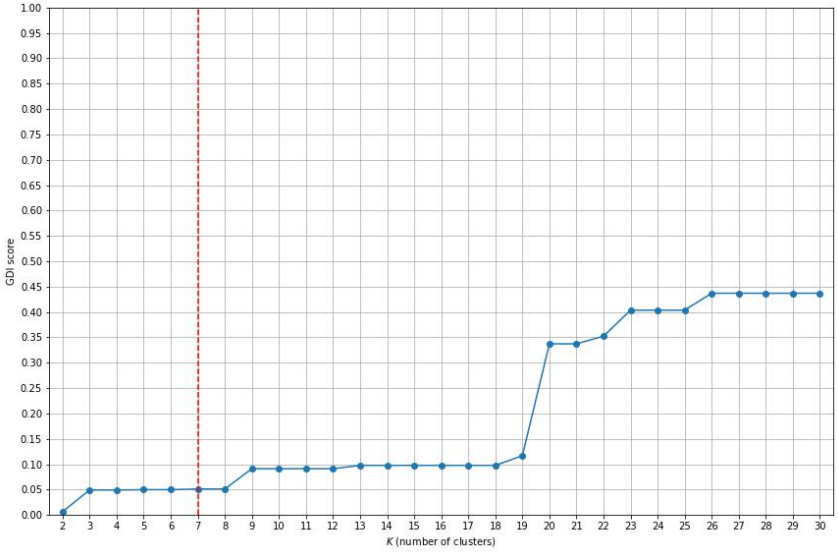


Figure 3.28: Evolution of the GDI score as a function of the number of K clusters.

As can be seen on the figure above, the GDI score of the hierarchical clustering with complete linkage for $K = 5$ is less than 0.1, which is very bad. Even if we decide to choose $K = 30$, this score would still not reach 0.5. We can therefore consider that this clustering method is unable to differentiate between female and male diagrams.

Hierarchical clustering with 'Ward' linkage

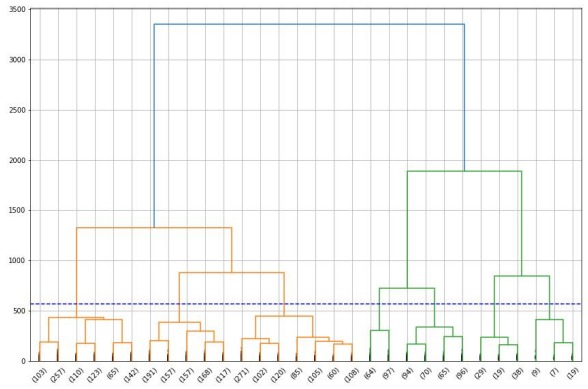


Figure 3.29: Dendrogram

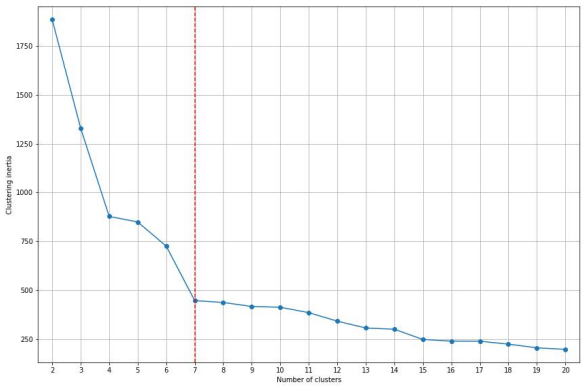


Figure 3.30: Elbow curve

According to the dendrogram and the elbow curve, the appropriate number of clusters to choose would be 5.

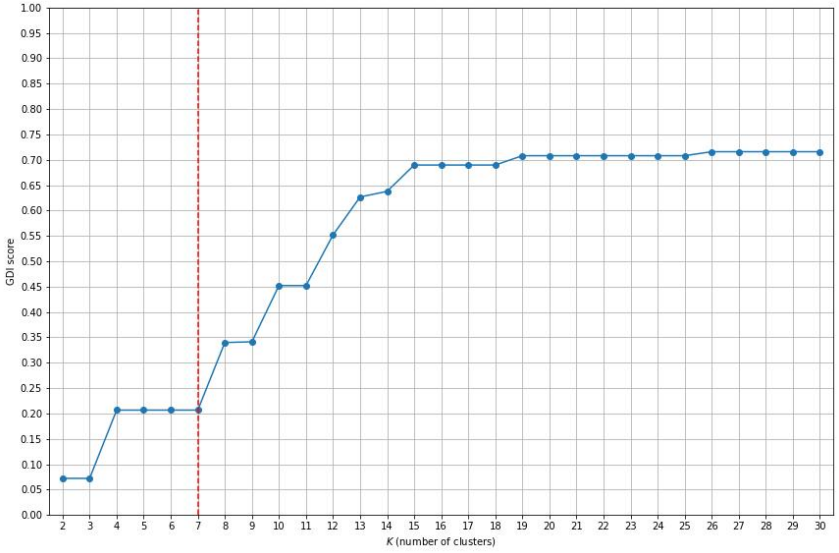


Figure 3.31: Evolution of the GDI score as a function of the number of K clusters.

As can be seen on the figure above, the GDI score of the hierarchical clustering with Ward linkage for $K = 5$ is barely above 0.1, which is still very bad. But it exceeds 0.5 from a number of clusters $K = 12$, which makes Ward linkage a better clustering method to differentiate male and female diagrams.

3.5.2 Silhouette vectors

Hierarchical clustering with Ward linkage

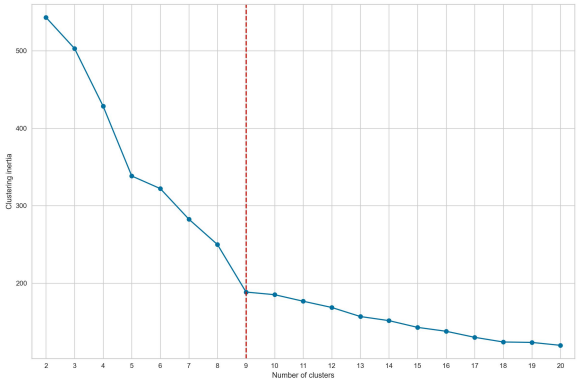


Figure 3.32: Distortion score elbow

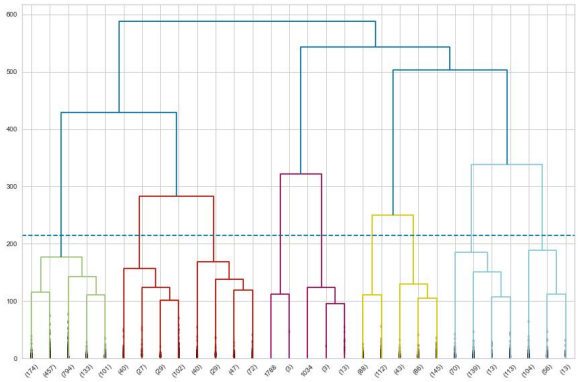


Figure 3.33: Dendrogram

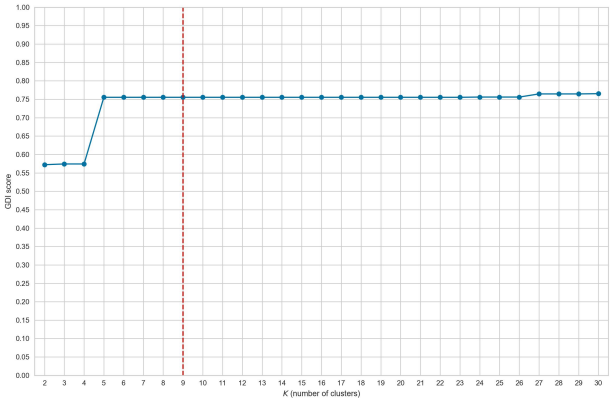


Figure 3.34: GDI Score

We observe that we have a decent score ($\simeq 0.572$) from 2 clusters and that from 5 clusters onwards, we obtain a very good GDI score ($\simeq 0.755$) and that this value varies very little when we increase the number of clusters.

K-Means

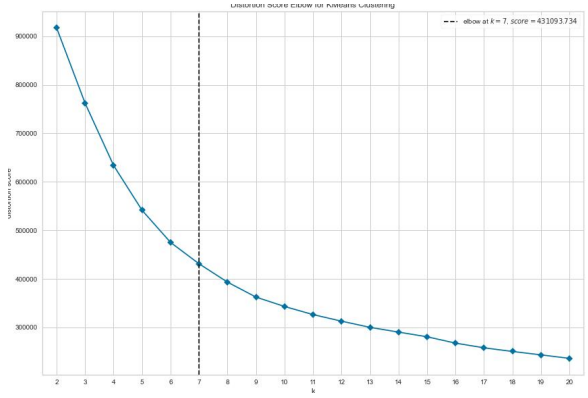


Figure 3.35: Distortion score elbow

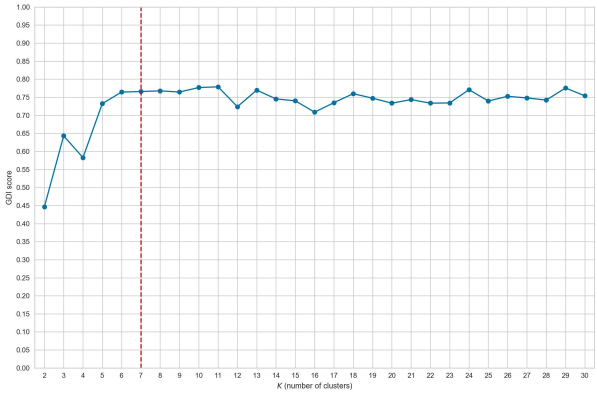


Figure 3.36: GDI Score

We observe that for 5 clusters we obtain a very good GDI score and that it remains good thereafter, even if we observe variations due to the randomisation of the algorithm.

K-Medoids (PAM)

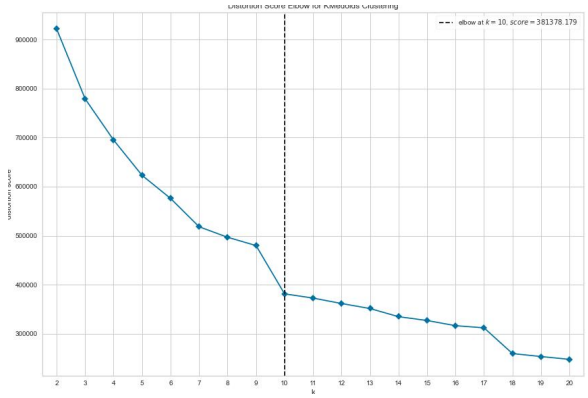


Figure 3.37: Distortion score elbow

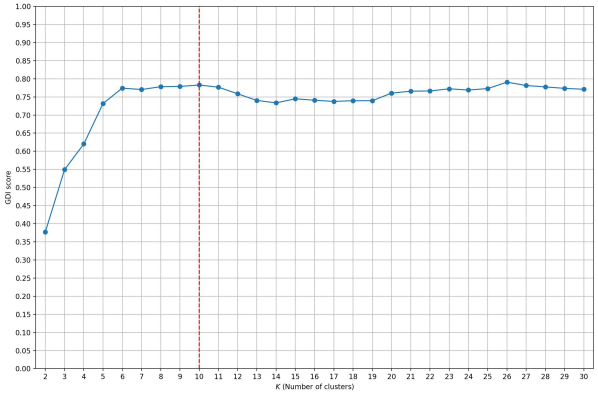


Figure 3.38: GDI Score

As for the K-Means clustering, we obtain a very good GDI score ($\simeq 0.732$) for 5 clusters and it remains good beyond (> 0.733), despite a small drop between 12 and 23 clusters.

3.5.3 First overview

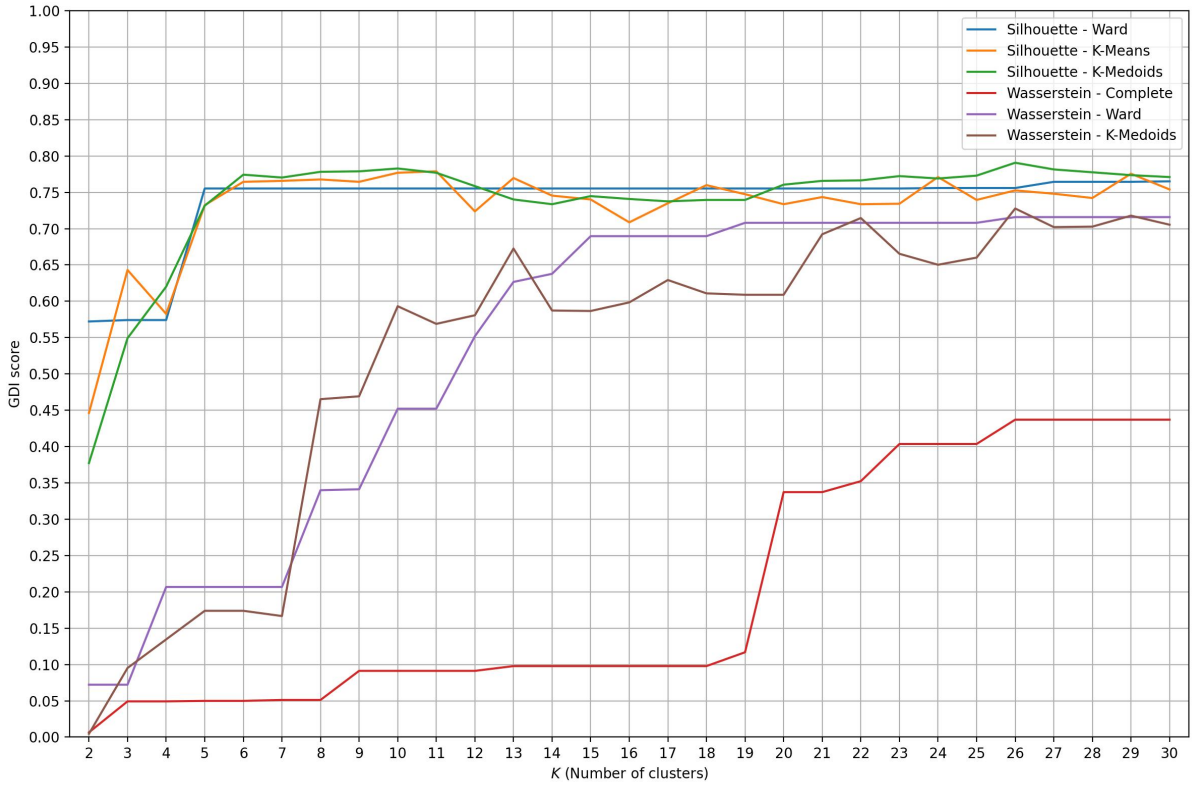


Figure 3.39: GDI Score - Overview

As can be seen in fig. 3.39, we get much better GDI scores using the vectors obtained from the persistence silhouettes than using the Wasserstein distances between persistence diagrams.

Using hierarchical clustering with Ward linkage, a good score is obtained even with few clusters and the score never decreases when the number of clusters is increased, so this seems to be the method of choice.

3.5.4 Restriction to trunks

When constructing the silhouettes, we used a weighting that tended to favour the H2 homologies corresponding to the trunks of the subjects, so the question then arises as to whether we would obtain better results by using only the points corresponding to the trunk of the body.

Extraction of points

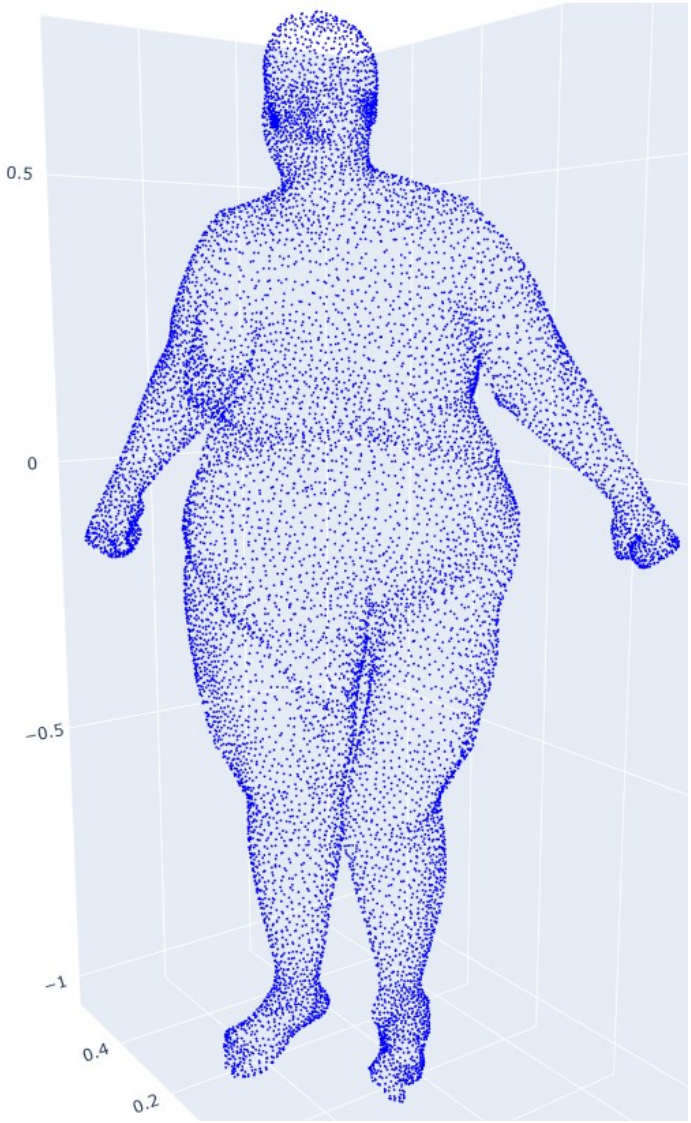


Figure 3.40: S_{init}

To extract the points corresponding to the trunk in an automated way, we proceed in several steps:

1. First of all, as the scans are off-axis, we need to rotate them by 53.13°.

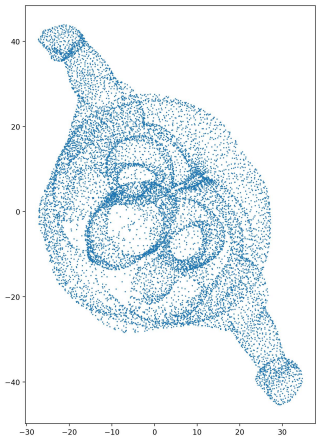


Figure 3.41: S_{init} - Top View

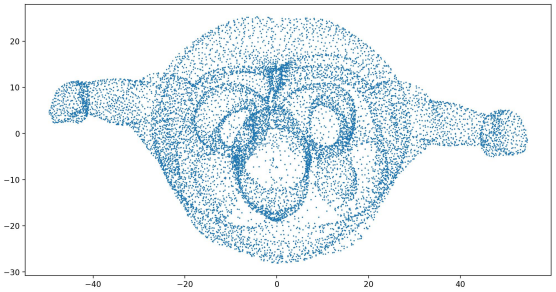


Figure 3.42: S_1 - Top View

Let S_{init} be the set of scan points of an individual and (x_1, x_2, x_3) the coordinates of a point

$$S_1 = \{(0.6x_1 - 0.8x_2, 0.8x_1 + 0.6x_2, x3) \mid (x_1, x_2, x_3) \in S_{init}\}$$

2. After normalising the scan to a height of 1.70m, we calculate l the average of the absolute values of the abscissa of the points in the height range of 66.5cm to 146.5cm.

$$S_2 = \{(x_1, x_2, x_3) \in S_1 \mid 66.5 \leqslant x_3 \leqslant 146.5\}$$

$$l = \frac{1}{s(S_2)} \sum_{(x_1,x_2,x_3)\in S_2} |x_1|$$

where $s(S_2)$ is the number of points in S_2.

3. Let S_3 and S_4 be such that

$$S_3 = \left\{(x_1, x_2, x_3) \in S_2 \;\middle|\; |x_1| < \frac{l'(4795 - 26x_3)}{2000} \text{ and } x_3 < 107.5\right\}$$

$$S_4 = \left\{(x_1, x_2, x_3) \in S_2 \;\middle|\; |x_1| < \frac{l'(2x_3 - 15)}{200} \text{ and } x_3 \geqslant 107.5\right\}$$

where $l' = \frac{111}{100}l$

Finally, we keep the set of points $S_{final} = S_3 \cup S_4$

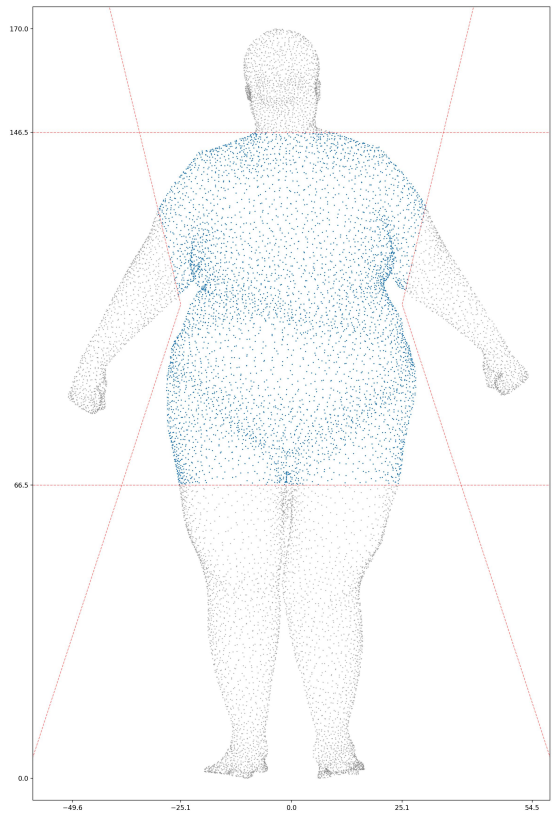


Figure 3.43: Trunk extraction - Front view

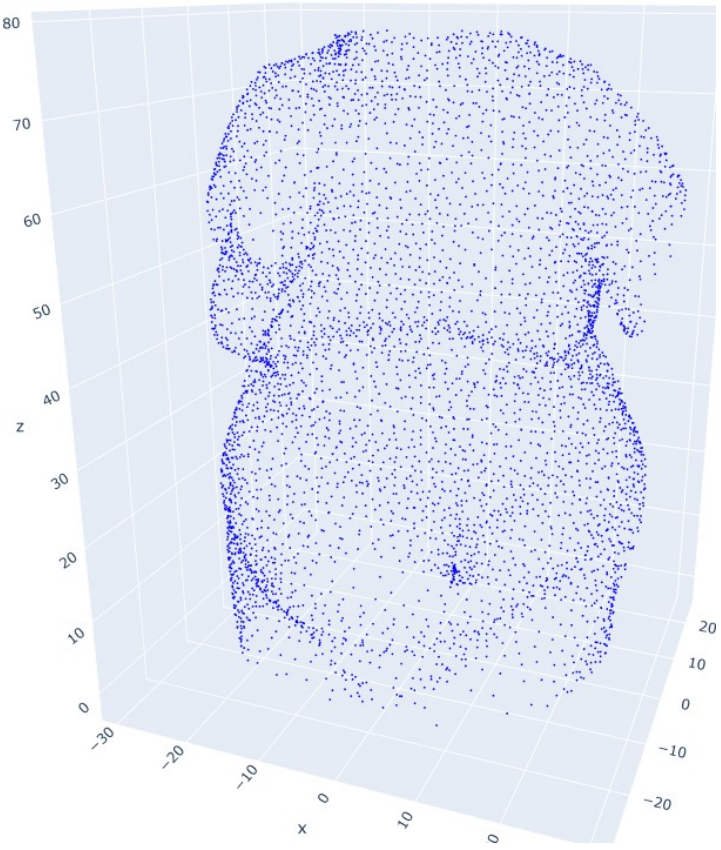


Figure 3.44: S_{final}

Trunks - Wasserstein distance

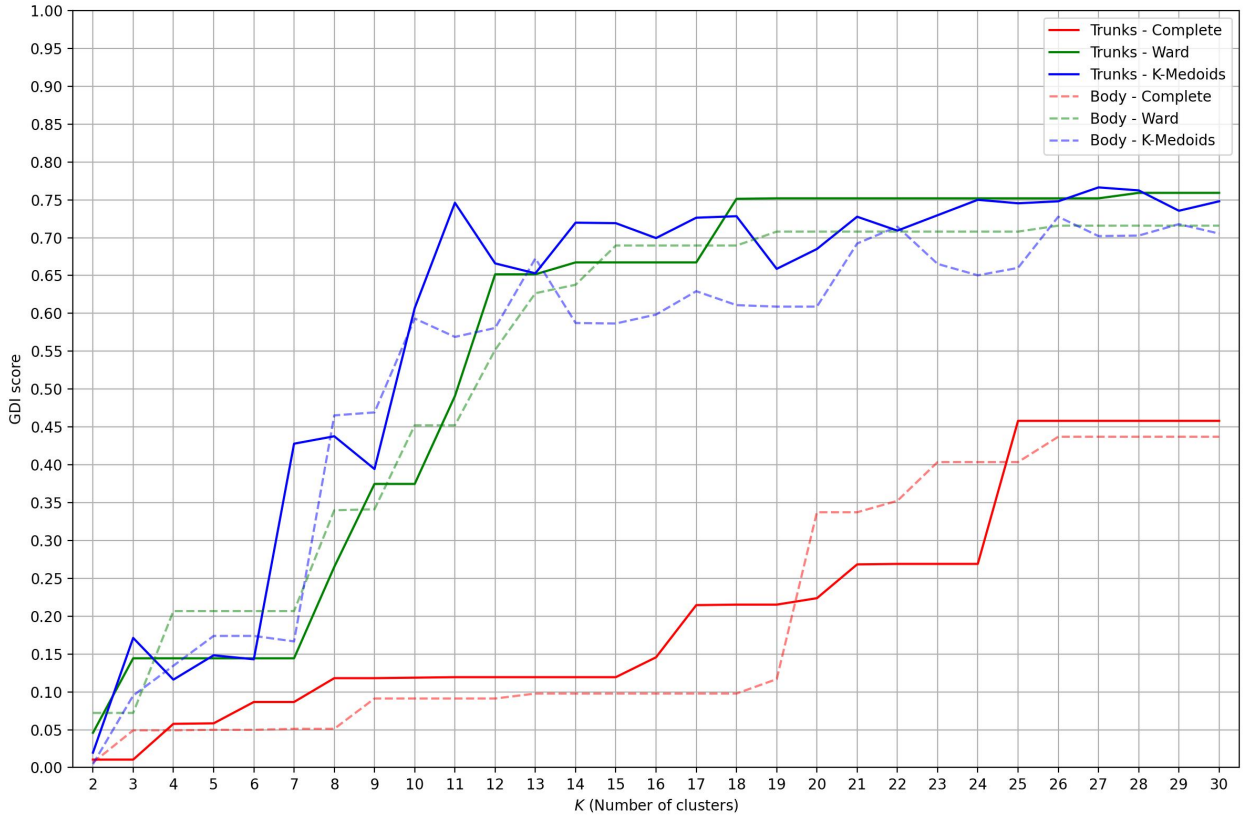


Figure 3.45: Comparison of whole body and trunk results

Table 3.4: Average GDI scores

	Complete	Ward	K-Medoids
Body	0.2	0.54	0.526
Trunk	0.21	0.553	0.582

Comparing the curves in fig. 3.45 and the average GDI scores (table 3.4), it appears that for clustering based on Wasserstein distances between persistence diagrams, it is not worth restricting to trunk points.

Trunks - Silhouette vectors

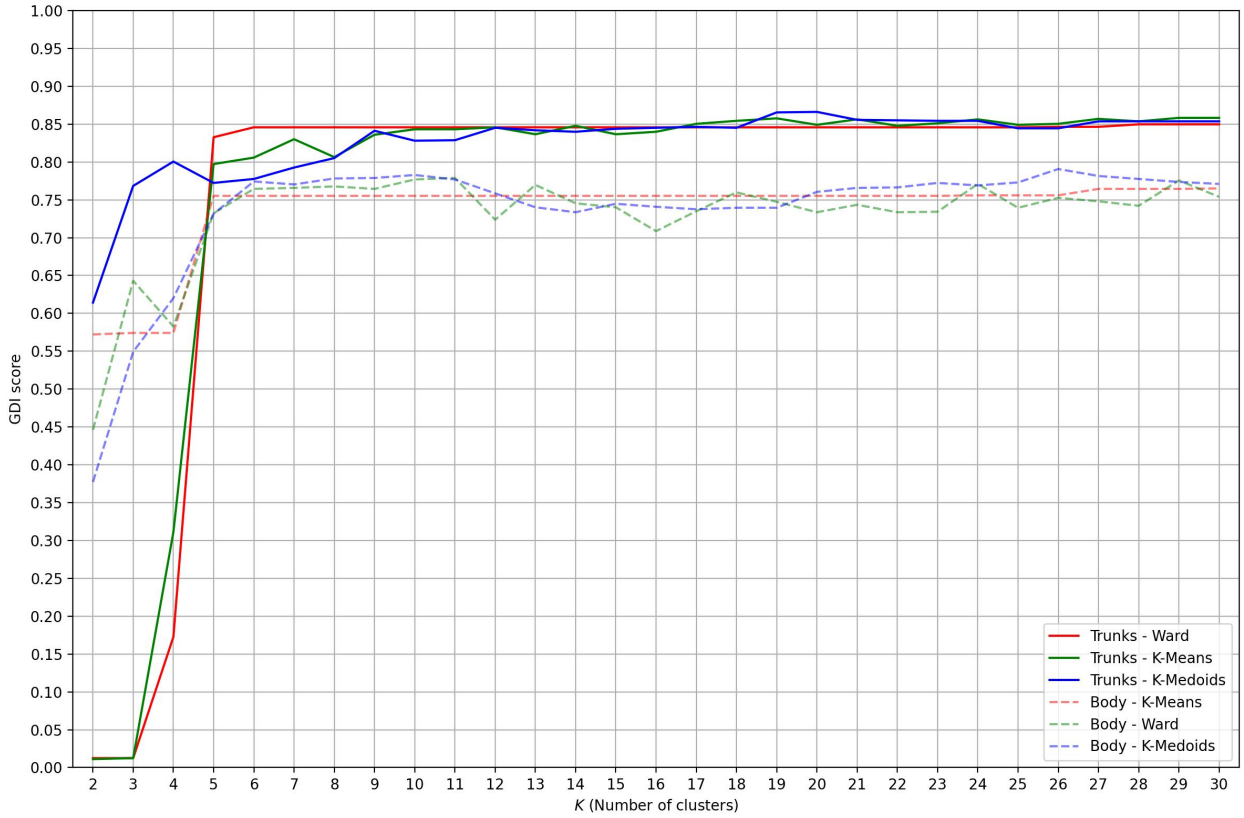


Figure 3.46: Comparison of whole body and trunk results

Table 3.5: Average GDI scores

	Ward	K-Means	K-Medoids
Body	0.738	0.73	0.737
Trunk	0.765	0.767	0.827

From the curves of fig. 3.46 and the average GDI scores (table 3.5), it appears that for clustering based on Silhouette persistence vectors, it is worth restricting to trunk points, especially for K-Medoids clustering.

3.5.5 Conclusion

Based on the GDI score to choose the clustering method, the K-Medoid clustering (PAM) applied to the trunk points seems to be the best.

3.6 Clustering results using Silhouette vectors

3.6.1 K-Medoid clustering applied to the trunk points

Male dataset

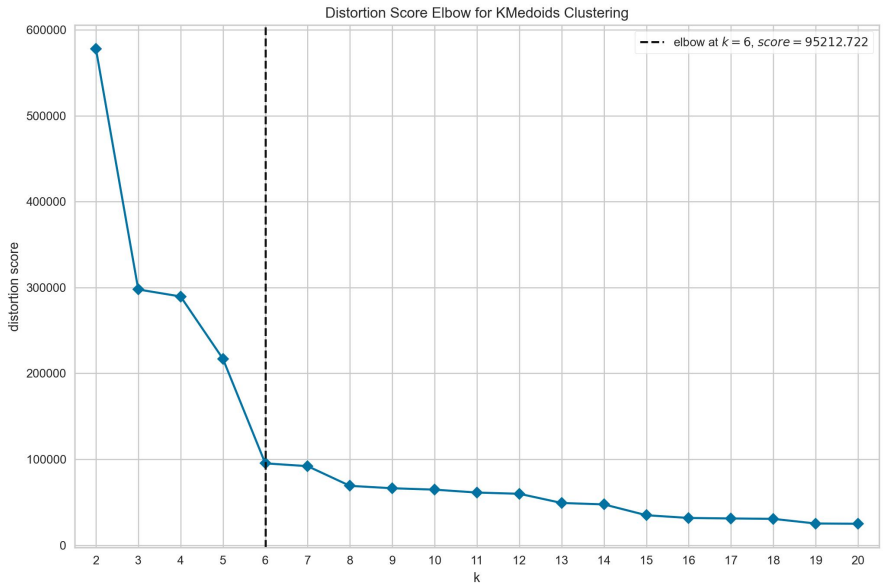


Figure 3.47: Elbow curve

According to the elbow curve, the optimal number of clusters seems to be 6.

Cluster n°	1	2	3	4	5	6
Size	22	9	3	618	441	424
Proportion	1.45%	0.59%	0.2%	40.74%	29.07%	27.95%

Table 3.6: Clustering results

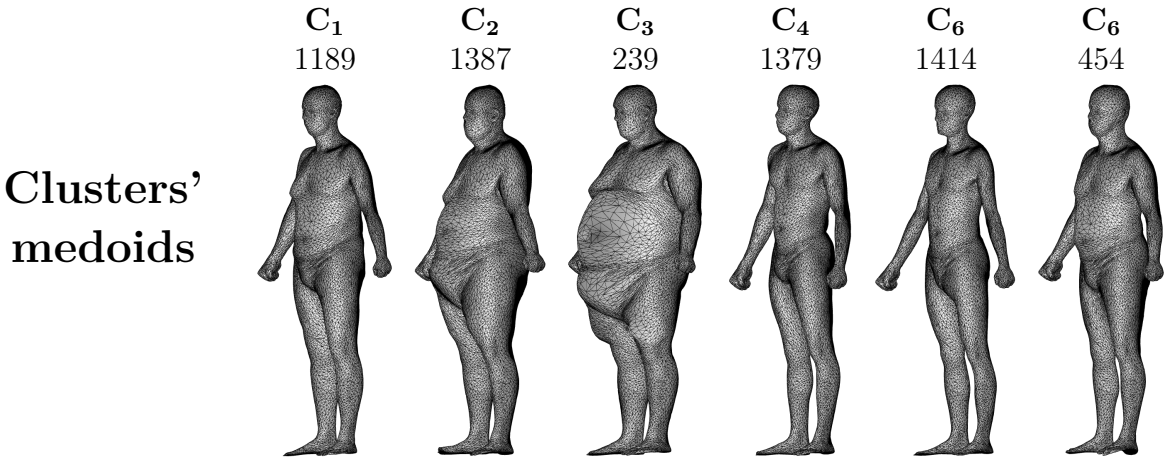


Figure 3.48: Indexes and scans of medoids.

Female dataset

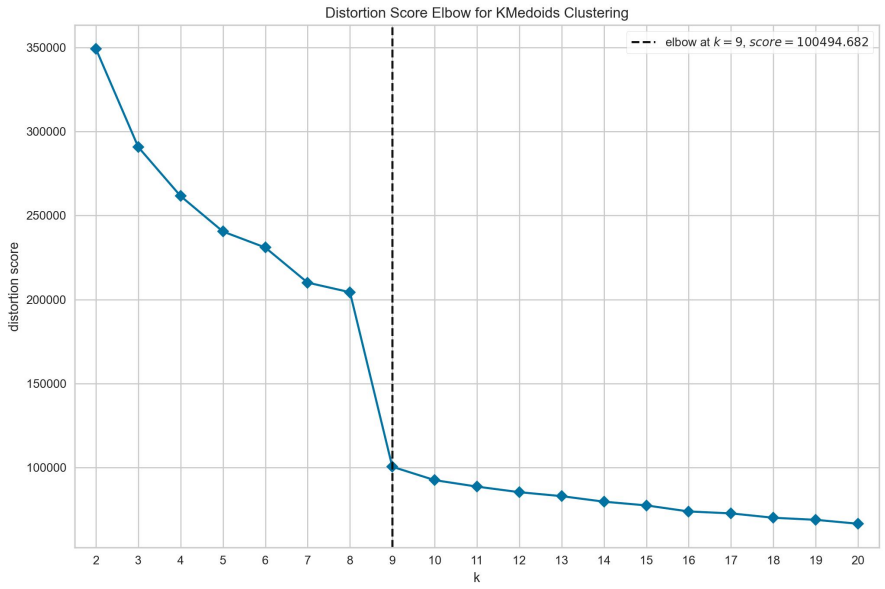


Figure 3.49: Elbow curve

According to the elbow curve, the optimal number of clusters seems to be 6.

Cluster n°	1	2	3	4	5	6	7	8	9
Size	253	148	196	154	115	2	287	144	232
Proportion	16.53%	9.67%	12.8%	10.06%	7.51%	0.13%	18.75%	9.41%	15.15%

Table 3.7: Clustering results

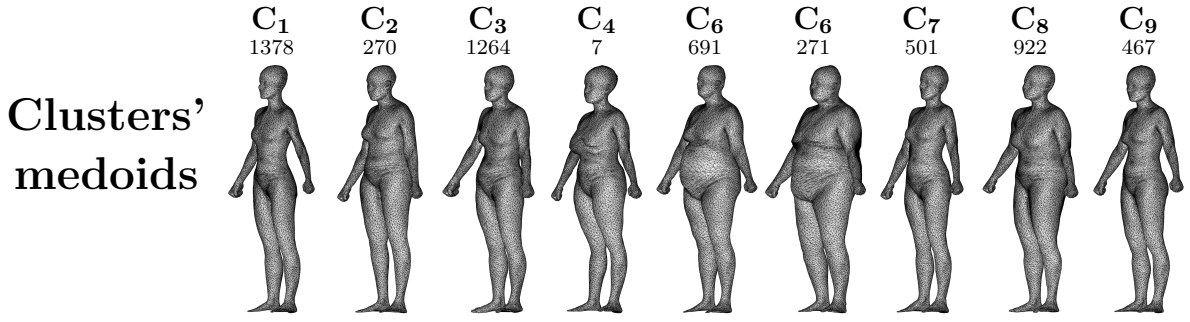


Figure 3.50: Indexes and scans of medoids.

Conclusion

We get more clusters with the female dataset and its individuals are much better distributed than for the male dataset. This is probably due to the fact that the weighting used to make the persistence silhouettes gives more weight to female characteristics.

4 Assessment of the internship

Having done my internship remotely, I was able to develop my autonomy. Once the project was presented, I worked freely and explored the avenues that seemed interesting to me to develop while carrying out the work I was asked to do as best I could.

Thanks to this internship, I was able to deepen my knowledge of unsupervised learning and learn new ones such as topological data analysis and Wasserstein distance.

Bibliography

[1] "Wikipedia: University of technology of troyes." https://en.wikipedia.org/.

[2] "Wikipedia: Institut français du textile et de l'habillement." https://fr.wikipedia.org/.

[3] R. O'Brien and W. C. Shelton, "Women's measurements for garments and pattern construction," *Miscellanous Publication*, no. 454, 1941.

[4] R. O'Brien, "An annotated list of literature references on arment sizes and body measurements," *D. S. Dept. AgT. MlSC. Pub.*, vol. 78, 1930.

[5] H. I. Douty, "Silhouette photography for the study of visual somatometry and body image," *The National Textiles and Clothing Meeting, Minneapolis, Minnesota*, 1968.

[6] J. Minott, *FFITing Commercial Patterns: The Minott Method*. Burgess Pub. Co, 1978.

[7] B. August and E. Count, *Complete Bonnie August Dress Thin System*. Macmillan Company, Incorporated, 1981.

[8] H. J. Armstrong, *Patternmaking for Fashion Design*. Harper & Row, 1987.

[9] K. Simmons, C. L. Istook, and P. Devarajan, "FEMALE FIGURE IDENTIFICATION TECHNIQUE (FFIT) FOR APPAREL PART II: DEVELOPMENT OF SHAPE SORTING SOFTWARE," vol. 4, no. 1, p. 15, 2004.

[10] K. Simmons, C. L. Istook, and P. Devarajan, "FEMALE FIGURE IDENTIFICATION TECHNIQUE (FFIT) FOR APPAREL PART I: DESCRIBING FEMALE SHAPES," vol. 4, no. 1, p. 16, 2004.

[11] S. L. Sokolowski and C. Bettencourt, "Modification of the Female Figure Identification Technique (FFIT) Formulas to Include Plus Size Bodies," in *Proceedings of 3DBODY.TECH 2020 - 11th International Conference and Exhibition on 3D Body Scanning and Processing Technologies, Online/Virtual, 17-18 November 2020*, (Online), Hometrica Consulting - Dr. Nicola D'Apuzzo, Nov. 2020.

[12] L. J. Connell, P. V. Ulrich, E. L. Brannon, M. Alexander, and A. Presley, "Body Shape Assessment Scale: Instrument Development Foranalyzing Female Figures," *Clothing and Textiles Research Journal*, no. 24, March 2006.

[13] K. Nakamura and T. Kurokawa, "Analysis and classification of three-dimensional trunk shape of women by using the human body shape model," *International Journal of Computer Applications in Technology*, vol. 34, no. 4, p. 278, 2009.

[14] F. S. Cottle, "Statistical Human Body Form Classification Methodology Development and Application," 2012.

[15] W. Park and S. Park, "Body shape analyses of large persons in South Korea," *Ergonomics*, vol. 56, pp. 692–706, Apr. 2013.

[16] F. Chazal and B. Michel, "An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists," *Frontiers in Artificial Intelligence*, vol. 4, p. 667963, Sept. 2021.

[17] R. Ghrist, *Elementary Applied Topology*. Createspace, 1st ed., 2014.

[18] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and S. Y. Oudot, "Proximity of Persistence Modules and Their Diagrams.," *Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry*, pp. 237 – 246, June 2009.

[19] F. Chazal, V. de Silva, M. Glisse, and S. Oudot, "The Structure and Stability of Persistence Modules," *SpringerBriefs in Mathematics*, 2016.

[20] Y. Yang, Y. Yu, Y. Zhou, S. Du, J. Davis, and R. Yang, "Semantic Parametric Reshaping of Human Body Models," in *2014 2nd International Conference on 3D Vision*, (Tokyo), pp. 41–48, IEEE, Dec. 2014.

[21] A. Taghribi, M. Canducci, M. Mastropietro, S. De Rijcke, K. Bunte, and P. Tiňo, "ASAP – A sub-sampling approach for preserving topological structures modeled with geodesic topographic mapping," *Neurocomputing*, vol. 470, pp. 376–388, Jan. 2022.

[22] J.-D. Boissonnat and C. Maria, "The Simplex Tree: an Efficient Data Structure for General Simplicial Complexes," *Algorithmica*, vol. 70, pp. 406–427, Nov. 2014. arXiv:2001.02581 [cs].

[23] G. N. Lance and W. T. Williams, "A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems," *The Computer Journal*, vol. 9, pp. 373–380, Feb. 1967.

[24] M. Pirashvili, L. Steinberg, F. Belchi Guillamon, M. Niranjan, J. G. Frey, and J. Brodzki, "Improved understanding of aqueous solubility modeling through topological data analysis," *Journal of Cheminformatics*, vol. 10, p. 54, Dec. 2018.