Application of Topological Data Analysis for Biomechanical Dataset

Sangman Jung

sangmanjung@khu.ac.kr

Supervisor: Prof. Kyungsoo Kim

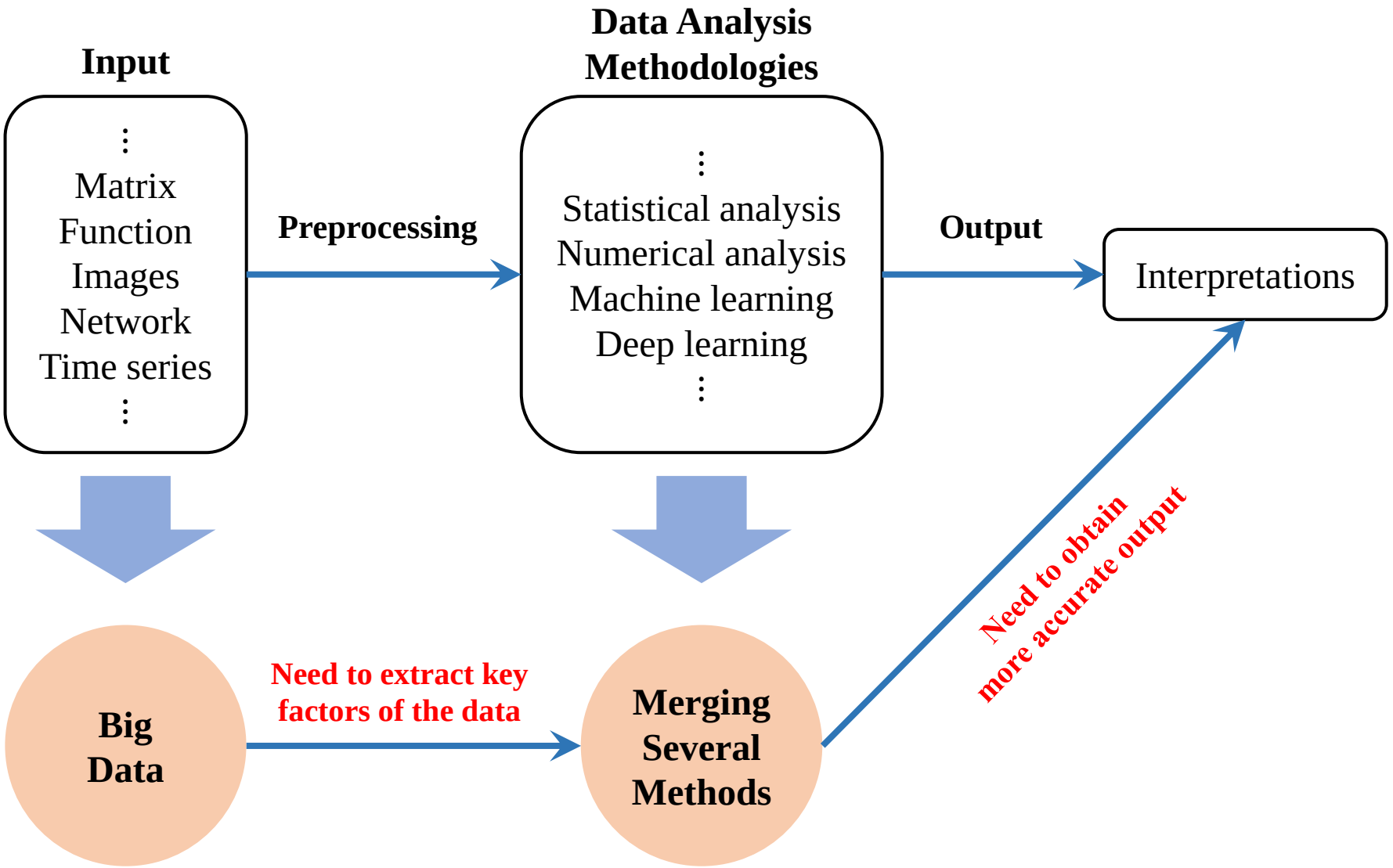
November 23, 2020

Department of Mathematics
Graduate School
Kyung Hee University

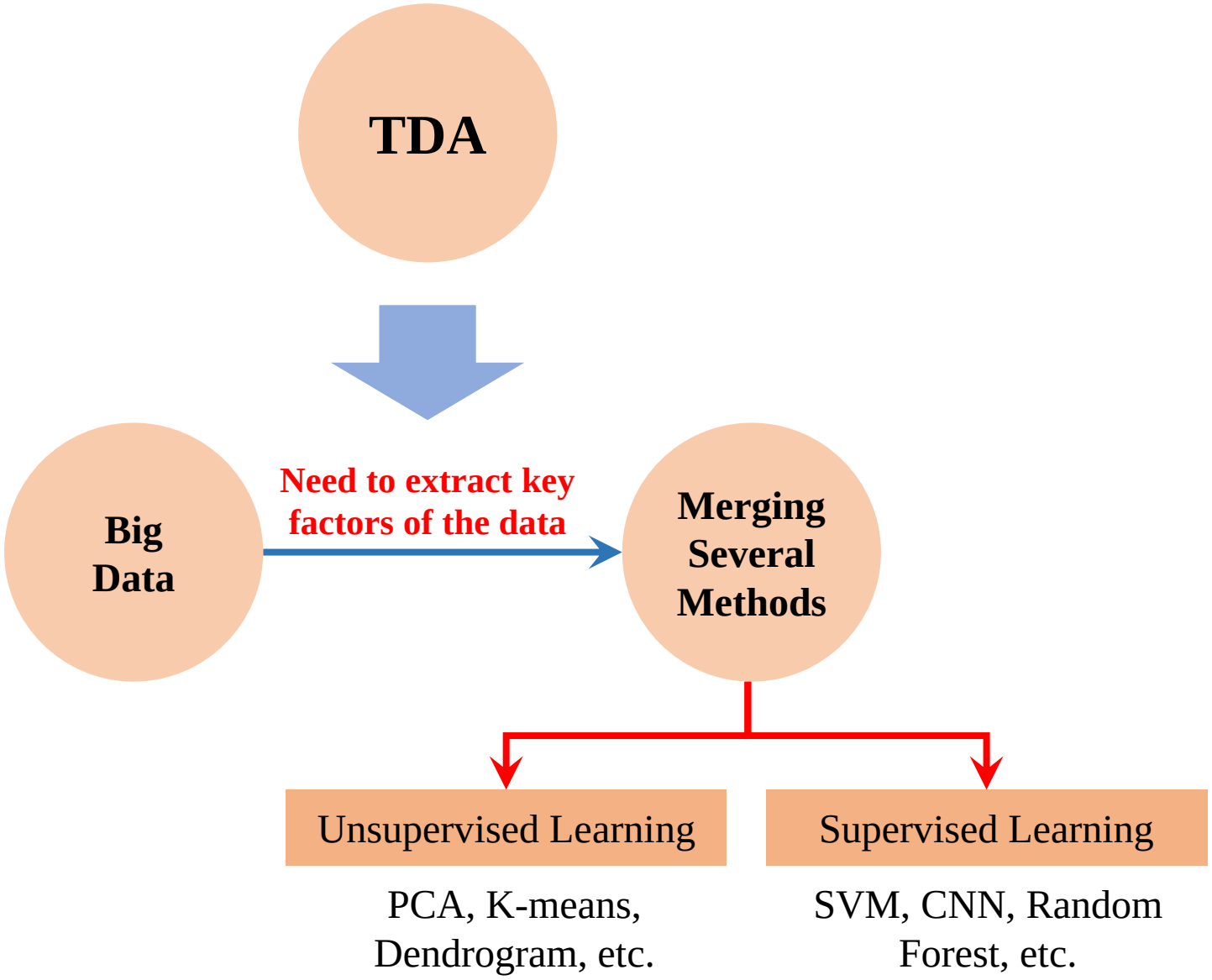
Table of Contents

Chapter 1. Introduction

1.1 Data analysis methodologies



1.2 Research motivation



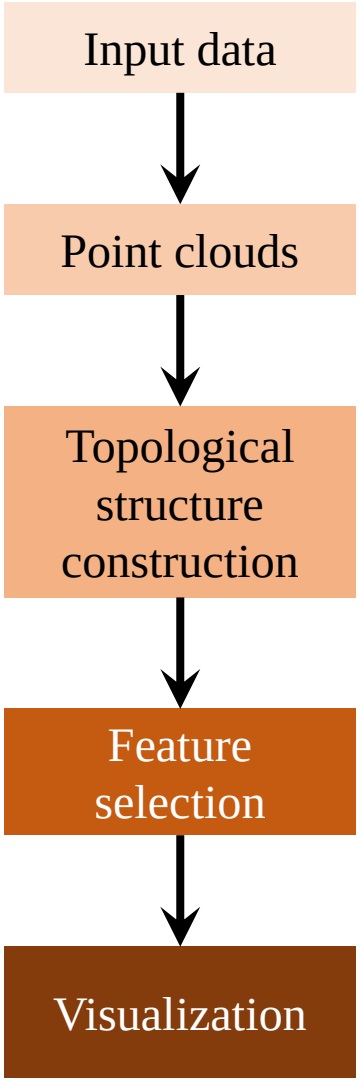
1.2 Research motivation

- **Why we use TDA**:
  - The recently proposed data analysis method
  - Extracts some topological features for the data
  - Effective in finding hidden features

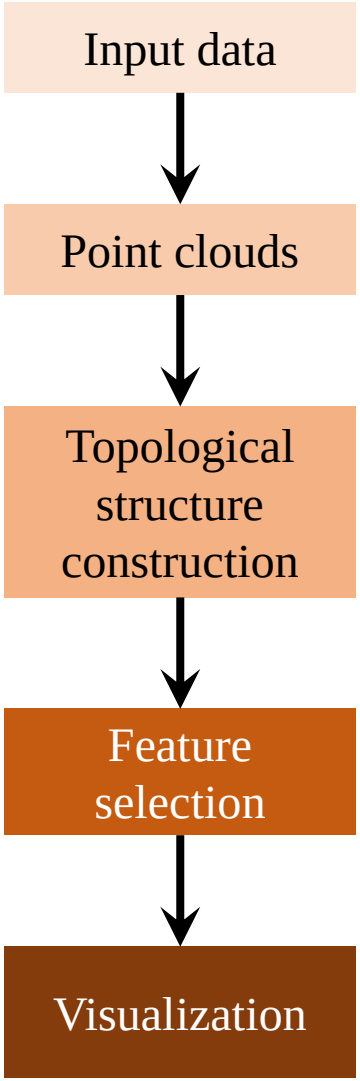
- **Purpose**:
  - To obtain the hidden features of each subject.
  - To quantify similarities or dissimilarities between subjects.

Chapter 2. Analysis Methods

2.1 Topological data analysis

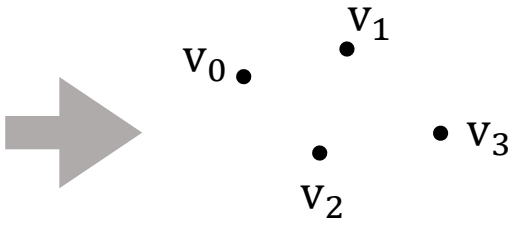


2.1 Topological data analysis



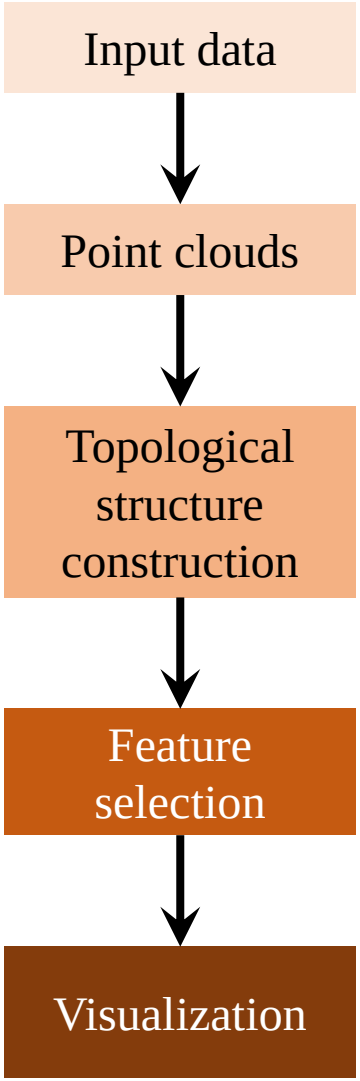
v_0	v_1	v_2	v_3
a_{11}	a_{12}	a_{13}	a_{14}
⋮	⋮	⋮	⋮
a_{n1}	a_{n2}	a_{n3}	a_{n4}

Input data



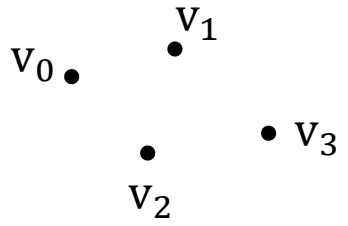
Point clouds

2.1 Topological data analysis

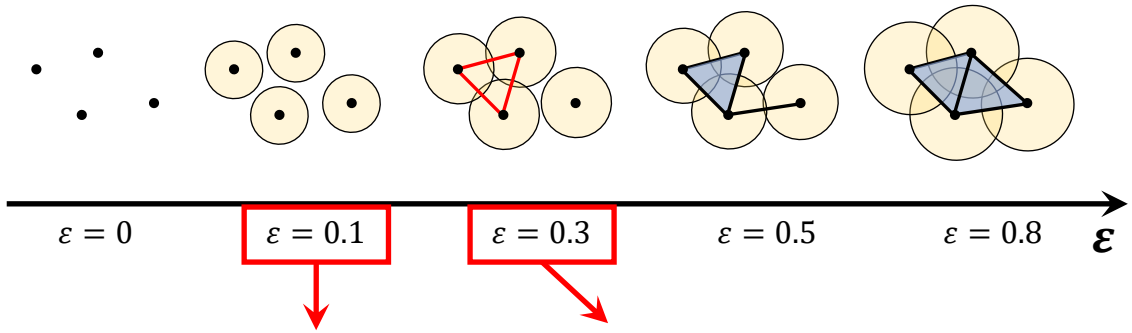


v_0	v_1	v_2	v_3
a_{11}	a_{12}	a_{13}	a_{14}
⋮	⋮	⋮	⋮
a_{n1}	a_{n2}	a_{n3}	a_{n4}

Input data



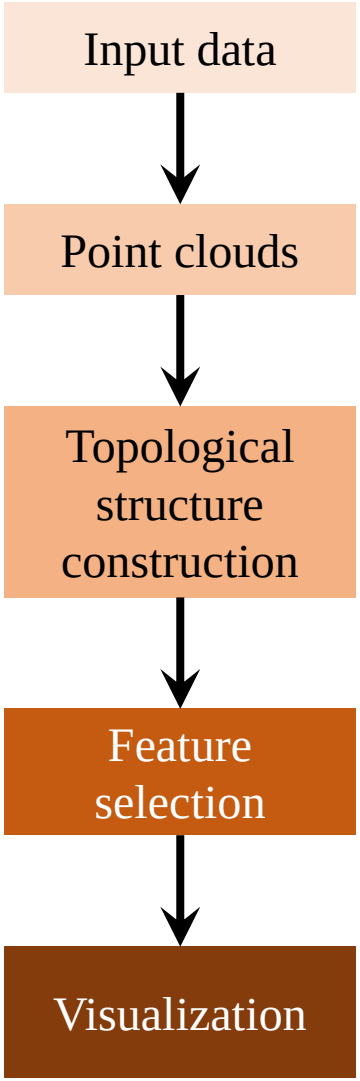
Point clouds



$\cdots\ (\beta_0, \beta_1, \beta_2) = (4, 0, 0) \quad (\beta_0, \beta_1, \beta_2) = (2, 1, 0)\ \cdots$

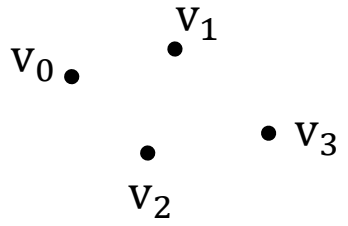
Topological structure construction & Feature selection

2.1 Topological data analysis

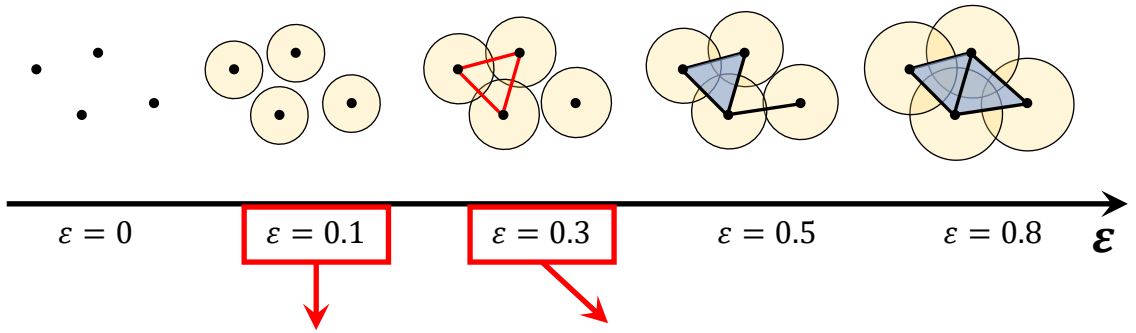


v_0	v_1	v_2	v_3
a_{11}	a_{12}	a_{13}	a_{14}
⋮	⋮	⋮	⋮
a_{n1}	a_{n2}	a_{n3}	a_{n4}

Input data

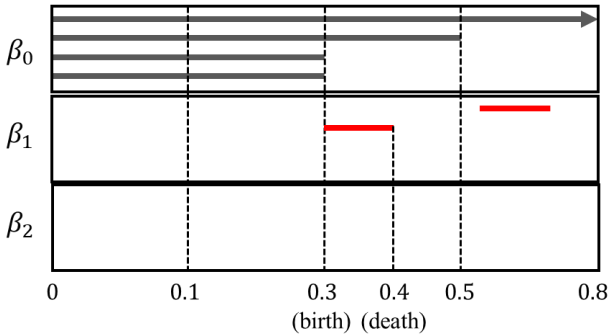


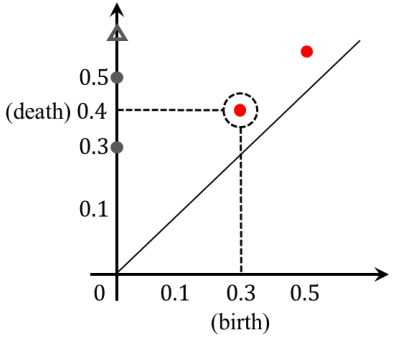
Point clouds



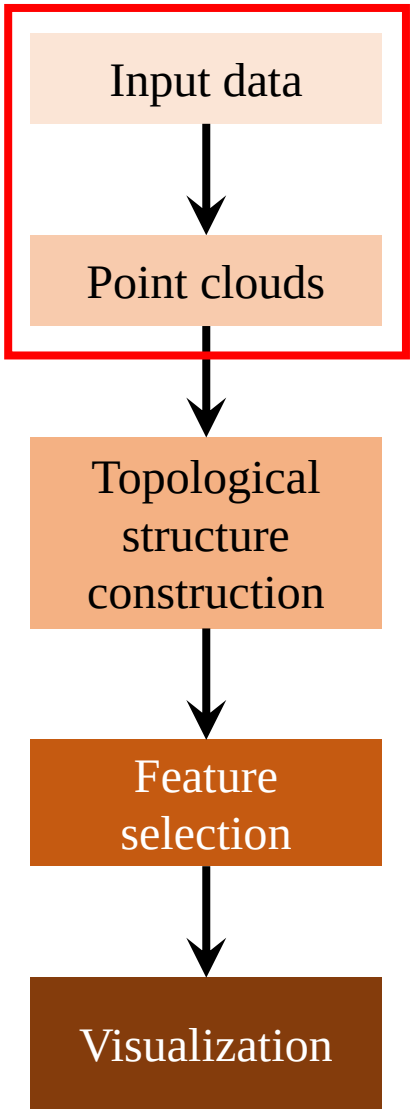
$\cdots\ (\beta_0, \beta_1, \beta_2) = (4, 0, 0) \quad (\beta_0, \beta_1, \beta_2) = (2, 1, 0)\ \cdots$

Topological structure construction & Feature selection





2.1 Topological data analysis



- **Point clouds construction:**

v_1	v_2	v_3
a_{12}	a_{13}	a_{14}
⋮	⋮	⋮
a_{n2}	a_{n3}	a_{n4}

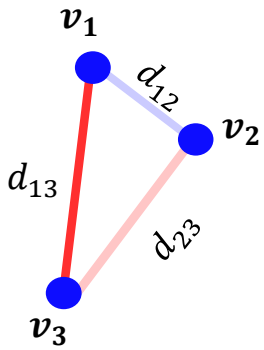
Input data

Compute the distance d_{ij} (e.g. Euclidean distance)

$$d_{ij} = d_{ji}$$

	v_1	v_2	v_3
v_1	0	d_{12}	d_{13}
v_2	d_{21}	0	d_{23}
v_3	d_{31}	d_{32}	0

Distance matrix (Point clouds)



Graph representation

	v_1	v_2	v_3
v_1	0	d_{12}	d_{13}
v_2	d_{21}	0	d_{23}
v_3	d_{31}	d_{32}	0

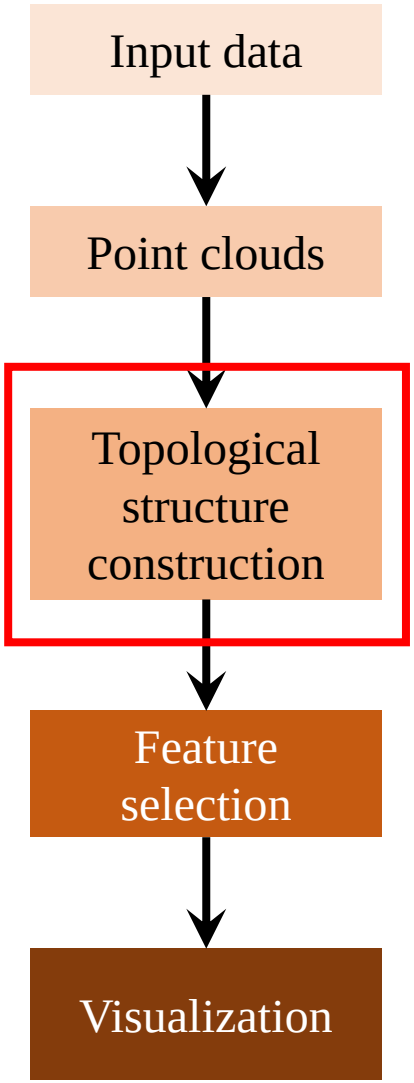
+

-

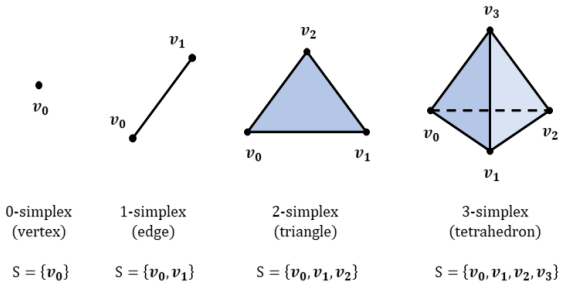
0

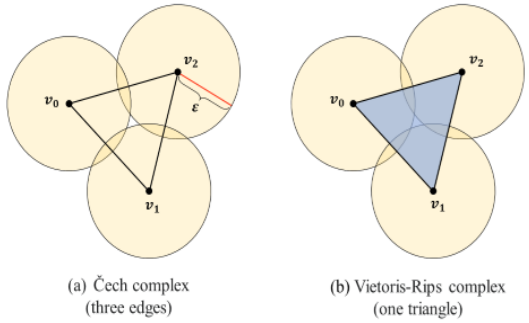
Heatmap representation

2.1 Topological data analysis

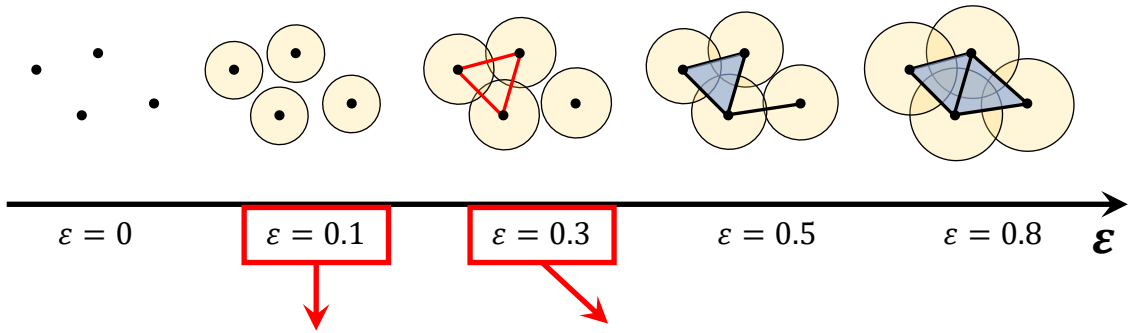


- **Topological structure construction:**



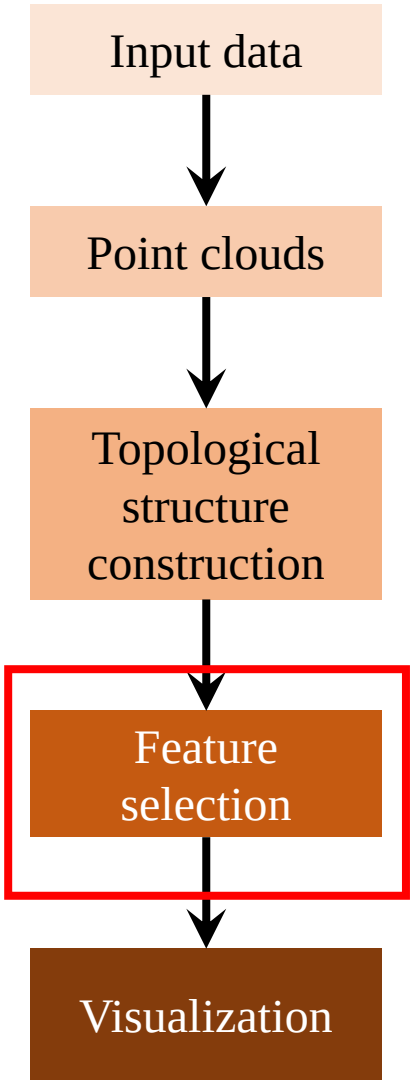


Topological structure construction

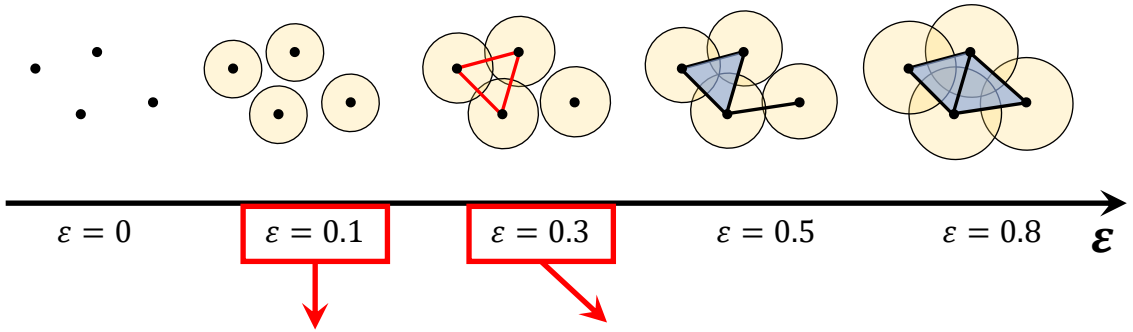


Feature selection

2.1 Topological data analysis



- **Feature selection:**



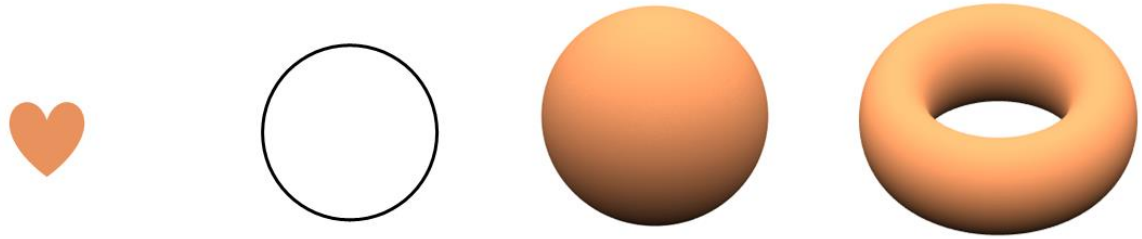
Betti Number $\boldsymbol{\beta_k}$:

(intuitively)

$\boldsymbol{\beta_0}$: the number of **connected components** (connectivity)

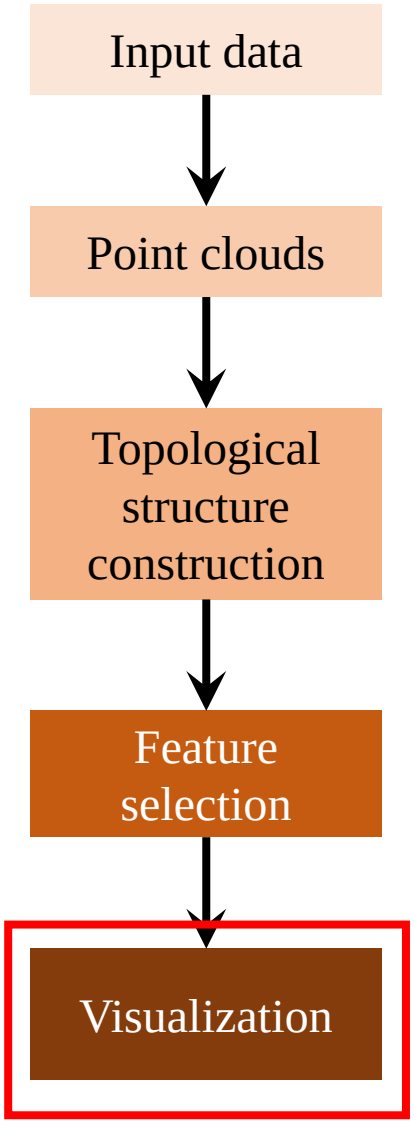
$\boldsymbol{\beta_1}$: the number of 1-dimensional **holes** or **loops**

$\boldsymbol{\beta_2}$: the number of enclosed solid **voids** (2-dimensional voids)

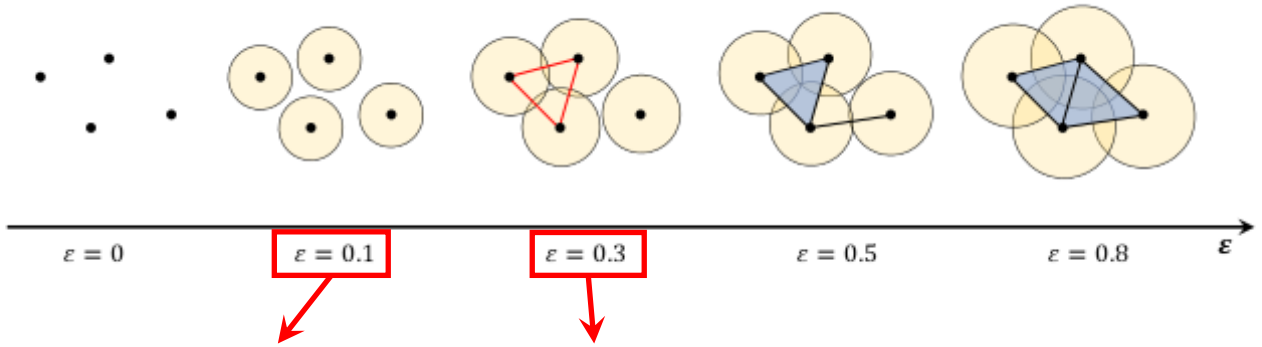


$(\beta_0, \beta_1, \beta_2) = (1, 0, 0)$ $(\beta_0, \beta_1, \beta_2) = (1, 1, 0)$ $(\beta_0, \beta_1, \beta_2) = (1, 1, 1)$ $(\beta_0, \beta_1, \beta_2) = (1, 2, 1)$

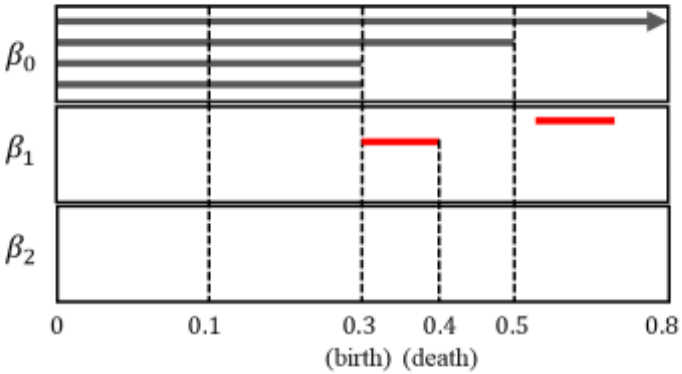
2.1 Topological data analysis



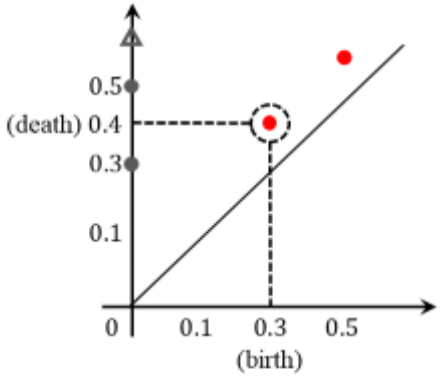
- **Visualization:**



$(\beta_0, \beta_1, \beta_2) = (4, 0, 0)$ $(\beta_0, \beta_1, \beta_2) = (2, 1, 0)$



Barcode



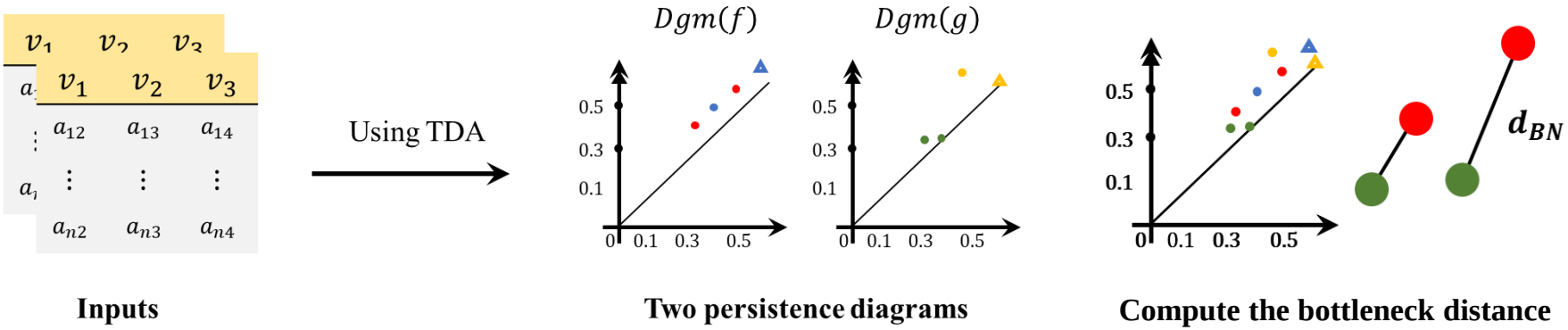
Persistence diagram

- **The case of multiple dataset:**

Compute the bottleneck distance between two persistence diagrams:

$$\boldsymbol{d_{BN}}\big(Dgm(f), Dgm(g)\big) = \inf_{\gamma} \sup_{1\le i\le m} \left\| x_i^f - \gamma\big(x_i^f\big) \right\|_\infty$$

$f, g : \mathbb{R} \to \mathbb{R}$, a bijection $\gamma : Dgm(f) \to Dgm(g)$ where $Dgm(f)$, $Dgm(g)$ are the persistence diagrams of f, g, respectively. Assume that $m = n$ where $|Dgm(f)| = m,\ |Dgm(g)| = n$.



Inputs **Two persistence diagrams** **Compute the bottleneck distance**

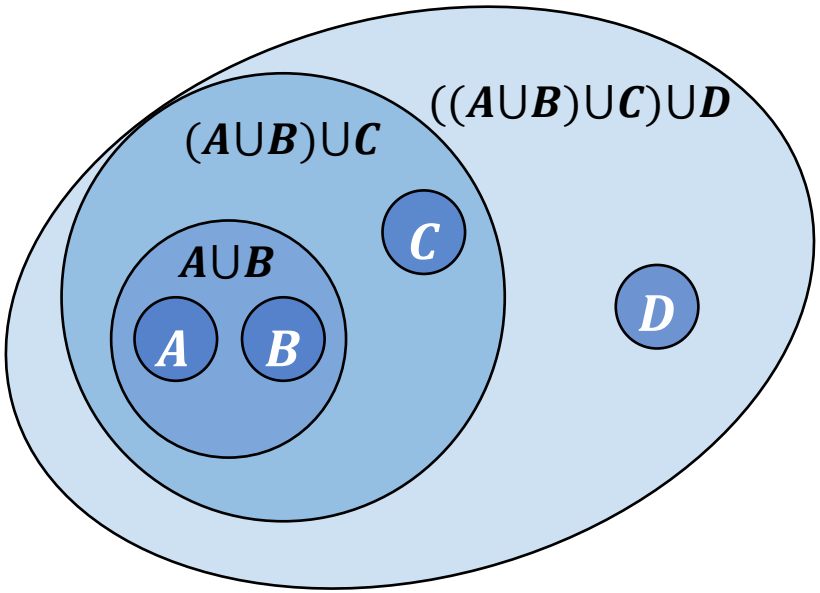
2.2 Conventional analysis methods

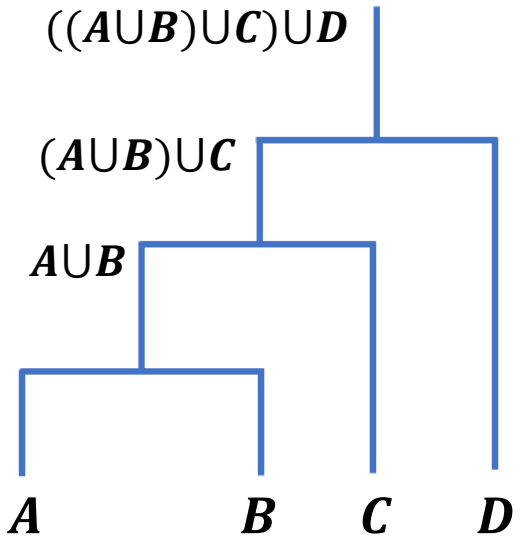
- **We used** the following methods with TDA:
  - The **single linkage dendrogram** in hierarchical clustering analysis
  - **Multidimensional scaling** for the dimensionality reduction

2.2 Conventional analysis methods

- **We used** the following methods with TDA:
  - **Single linkage dendrogram**
  - **Multidimensional scaling**

Dendrogram :
The diagram that **shows the hierarchical relationship** between objects such as **stem of a tree**.

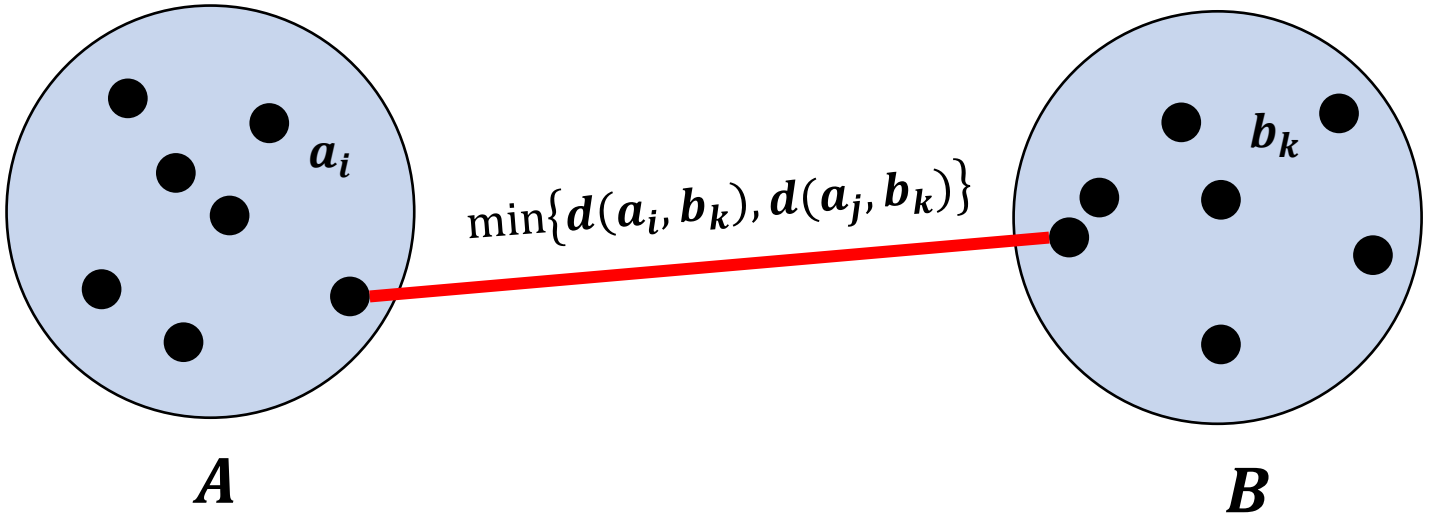




2.2 Conventional analysis methods

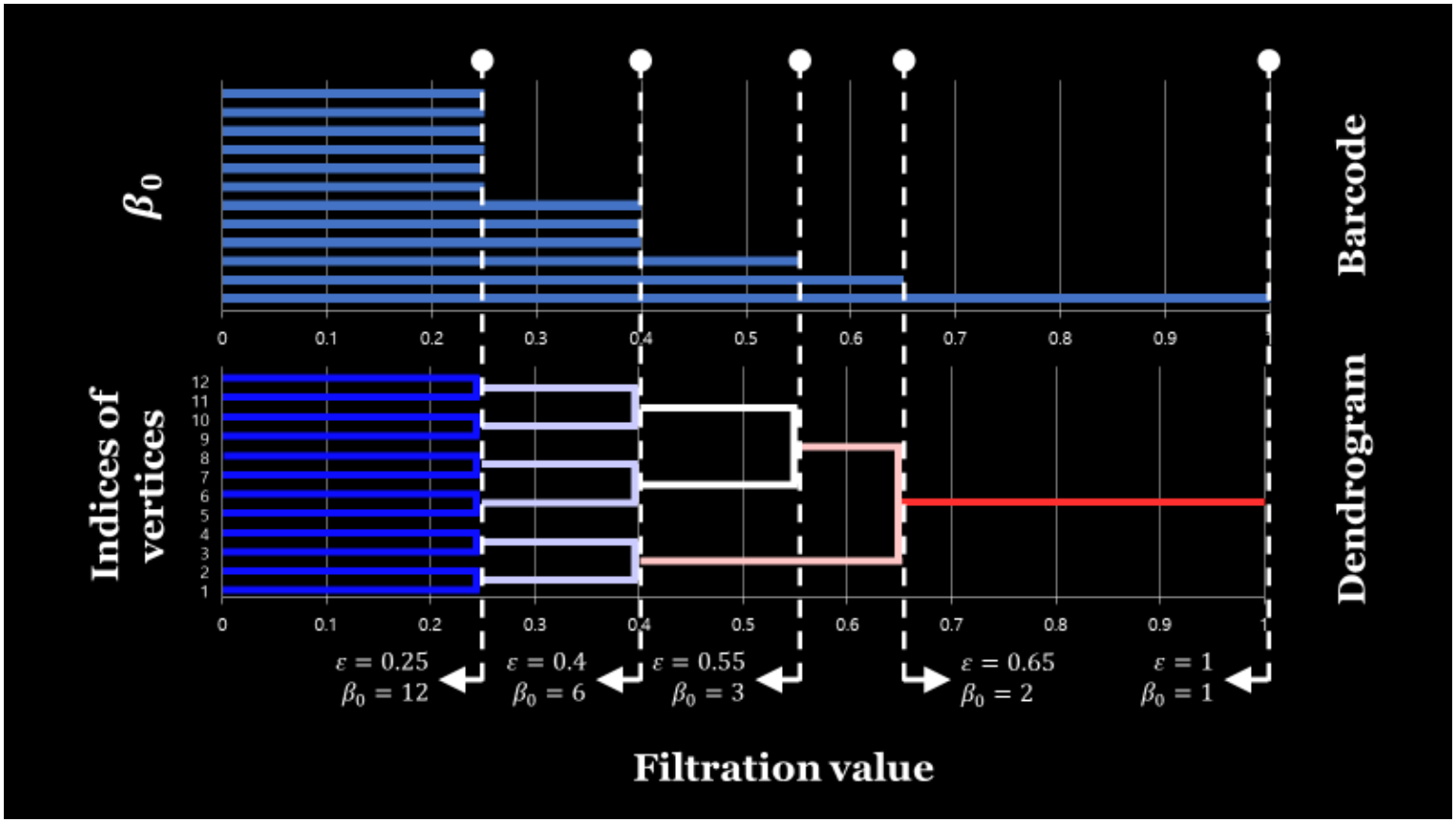
- **We used** the following methods with TDA:
  - **Single linkage dendrogram**
  - **Multidimensional scaling**

Single linkage method :
The linkage method that **merges** based on the **minimum** value of each cluster.



2.2 Conventional analysis methods

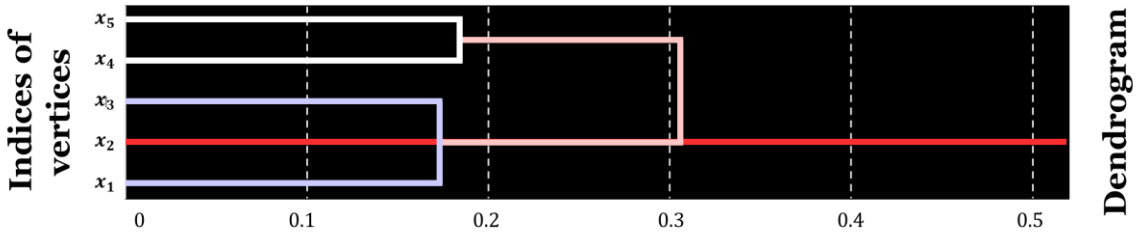
- **Relationship to barcode for β_0 :**
  - The β_0 barcode for the Vietoris-Rips construction is the same as the construction of the single linkage dendrogram [*].



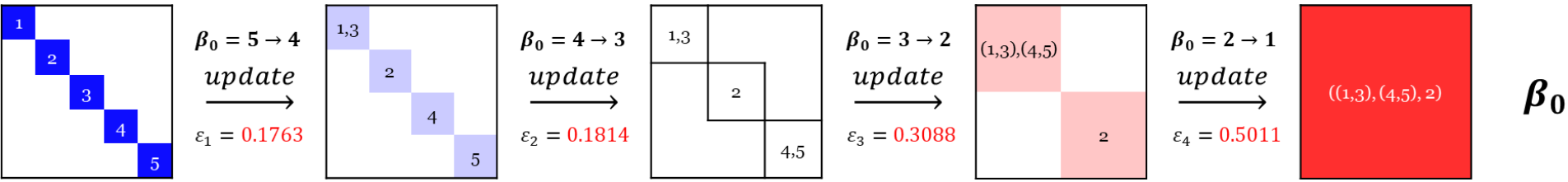
[*] Lee, H., Kang, H., Chung, M. K., Kim, B. N., & Lee, D. S. (2012). Persistent brain network homology from the perspective of dendrogram. IEEE transactions on medical imaging, 31(12), 2267-2277.

2.2 Conventional analysis methods

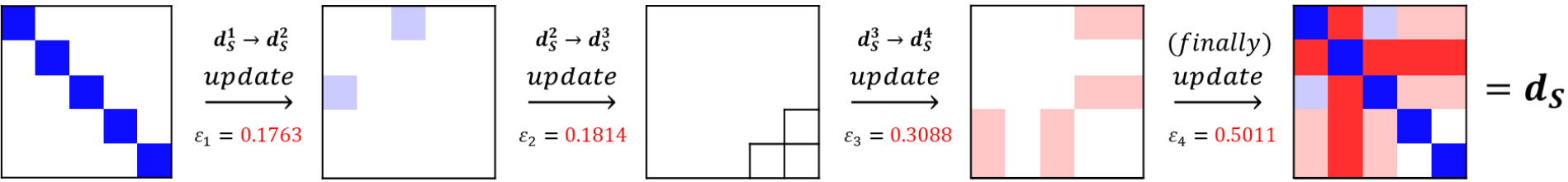
- **Practical use of the information of the dendrogram:**



- **Barcode for β_0 with clustering information:**



- **Single linkage dendrogram as a matrix form:**

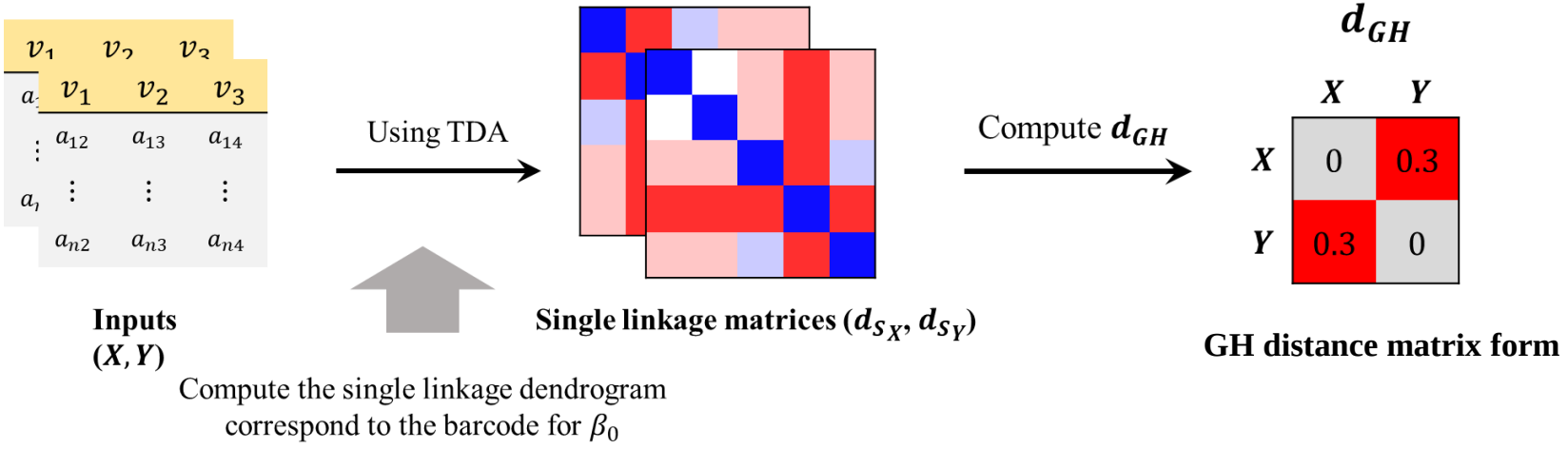


2.2 Conventional analysis methods

- **The case of multiple dataset:**

Compute the Gromov-Hausdorff distance (GH distance) between two single linkage matrices d_{S_X} and d_{S_Y}:

$$d_{GH}(X,Y) = \frac{1}{2}\max_{\forall i,j}\left|d_{S_X}((x_i,x_j) - d_{S_Y}(y_i,y_j)\right|$$



2.2 Conventional analysis methods

- **Multidimensional Scaling (MDS) :**

Input (Dissimilarity matrix) → **Use MDS** (Minimize the loss) → **Output** (MDS coordinates)

$$D = \begin{pmatrix} & V_1 & \cdots & V_p \\ & 0 & \cdots & d_{1p} \\ & \vdots & \ddots & \vdots \\ & d_{p1} & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{p \times p}$$

$d_{ij} = d_{ji}, i, j = 1, \cdots p,$

$V_1 \quad \cdots \quad V_p$ are variables.

Find q vectors $x_1, \cdots x_p \in \mathbb{R}^q$ such that

$$\|x_i - x_j\| \approx d_{ij}, \quad i, j = 1, \cdots p.$$

Criterions

⋮
strain (classical)
metricstress (metric)
stress (non-metric)
⋮

	Dim 1	Dim 2
V_1	x_{11}	x_{21}
	⋮	⋮
V_p	x_{1p}	x_{2p}

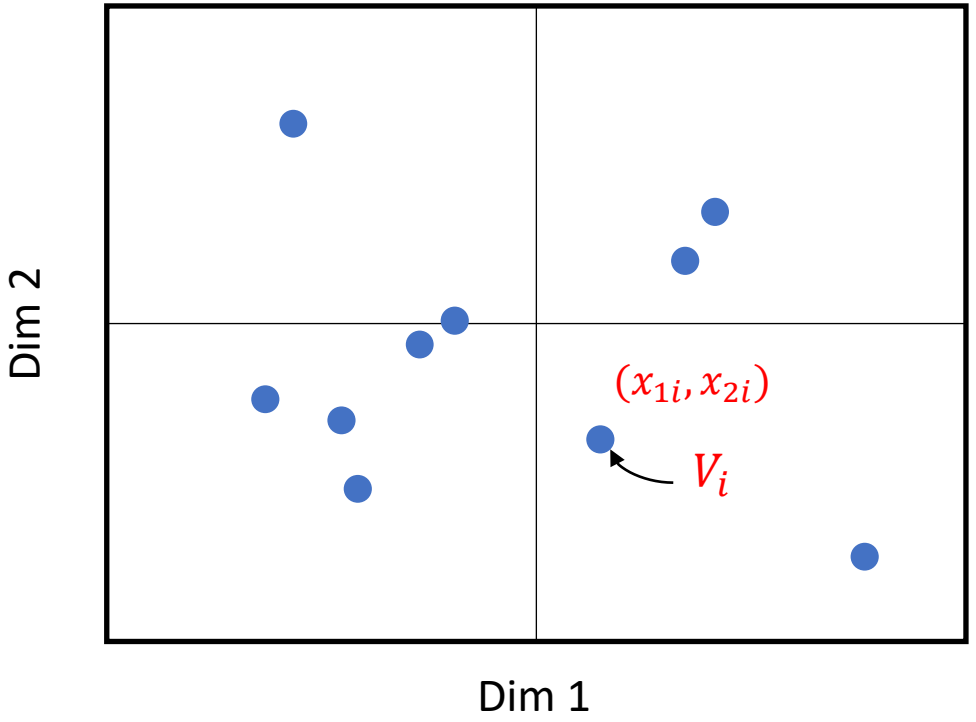
Obtain the q-dimensional coordinates.
In this example, $q = 2$.

2.2 Conventional analysis methods

- **Multidimensional Scaling (MDS)** :

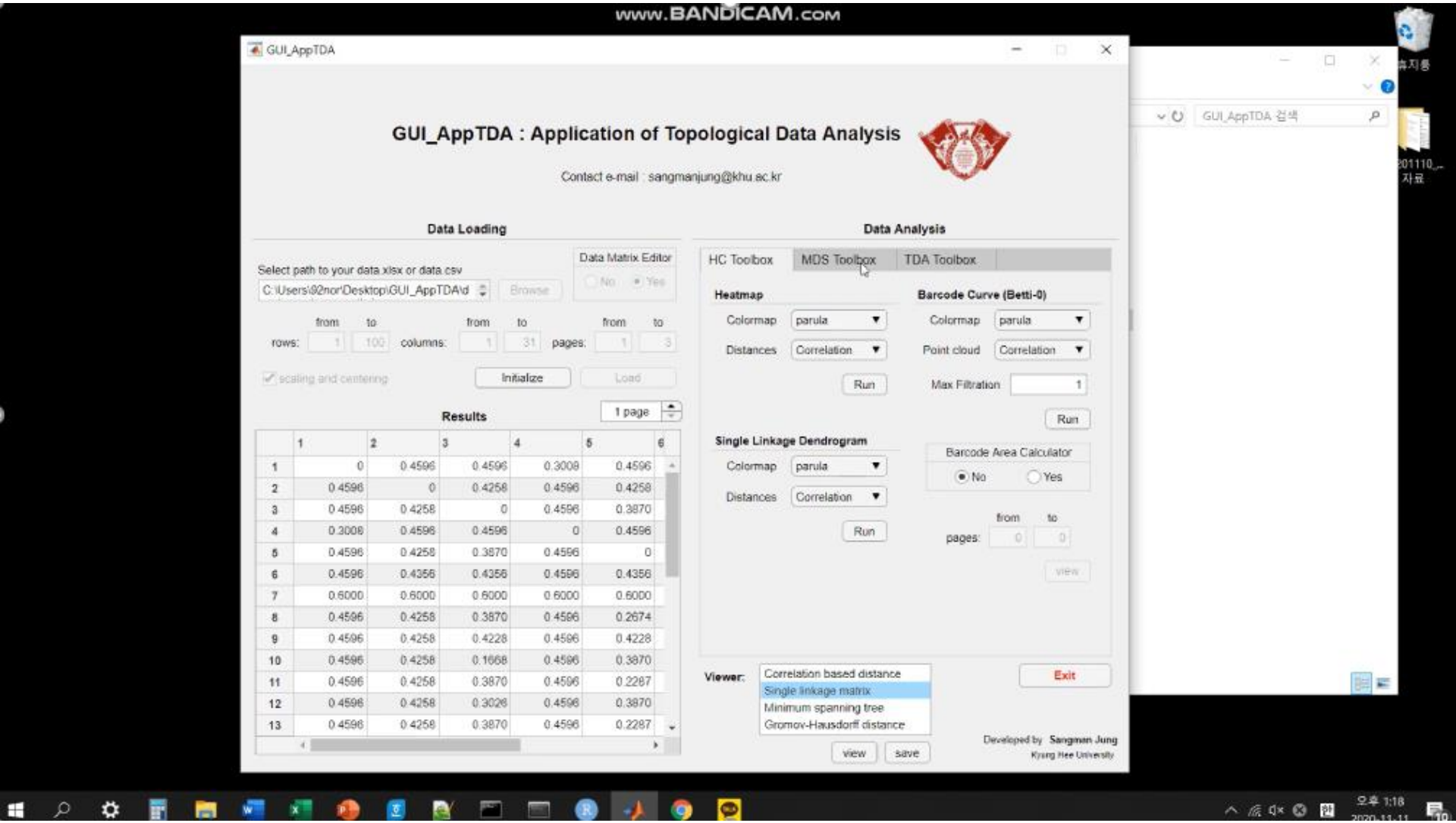
Output

	Dim 1	Dim 2
V_1	x_{11}	x_{21}
	⋮	⋮
V_p	x_{1p}	x_{2p}



Chapter 3. Development of GUI program

The video for the execution of GUI_AppTDA



Chapter 4. Application

4.1 Biomechanical Dataset

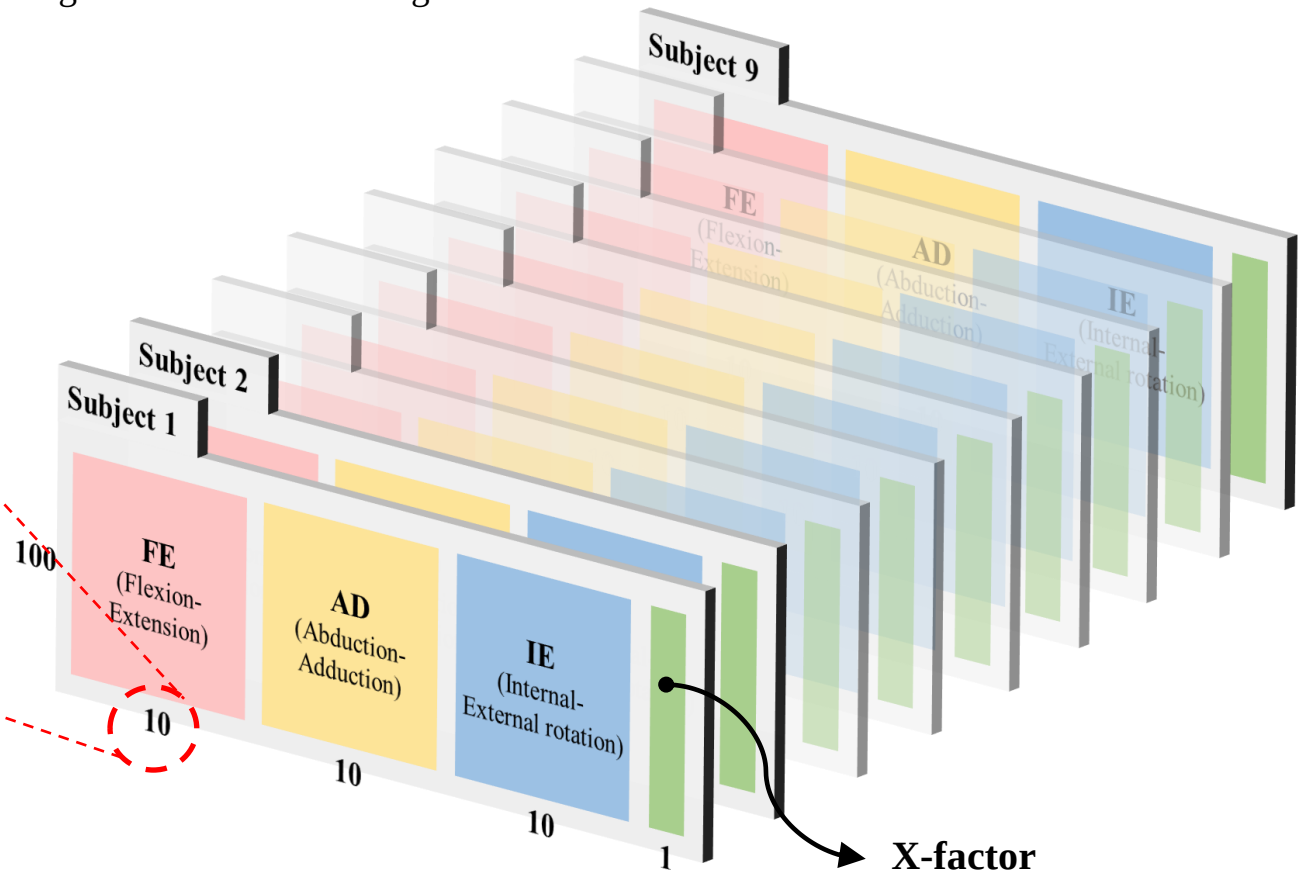
Golf swing motion
(it is normalized 100 sec)

Measure:

- **10 joint angles of 3 movements FE (굽힘/폄) , AD (벌림/모음) , IE (내/외부 회전)**
- **X-factor :** the difference between the rotation angles for the upper and lower bodies while the golfer takes a backswing

(Left + Right)
Shoulder, Elbow, Hip, Knee

Upper Trunk
Lower Trunk

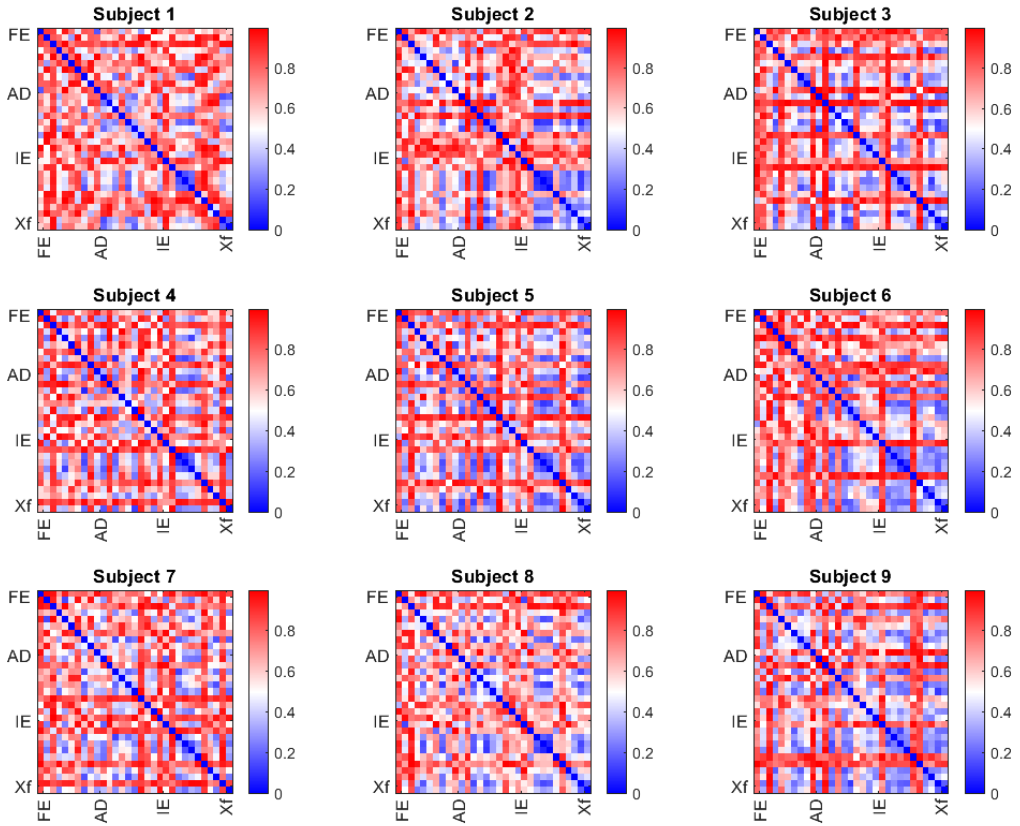


Data size:

$$\text{seconds} \times (\text{anlges} + \text{X-factor}) \times \text{subjects} = 100 \times (30 + 1) \times 9$$

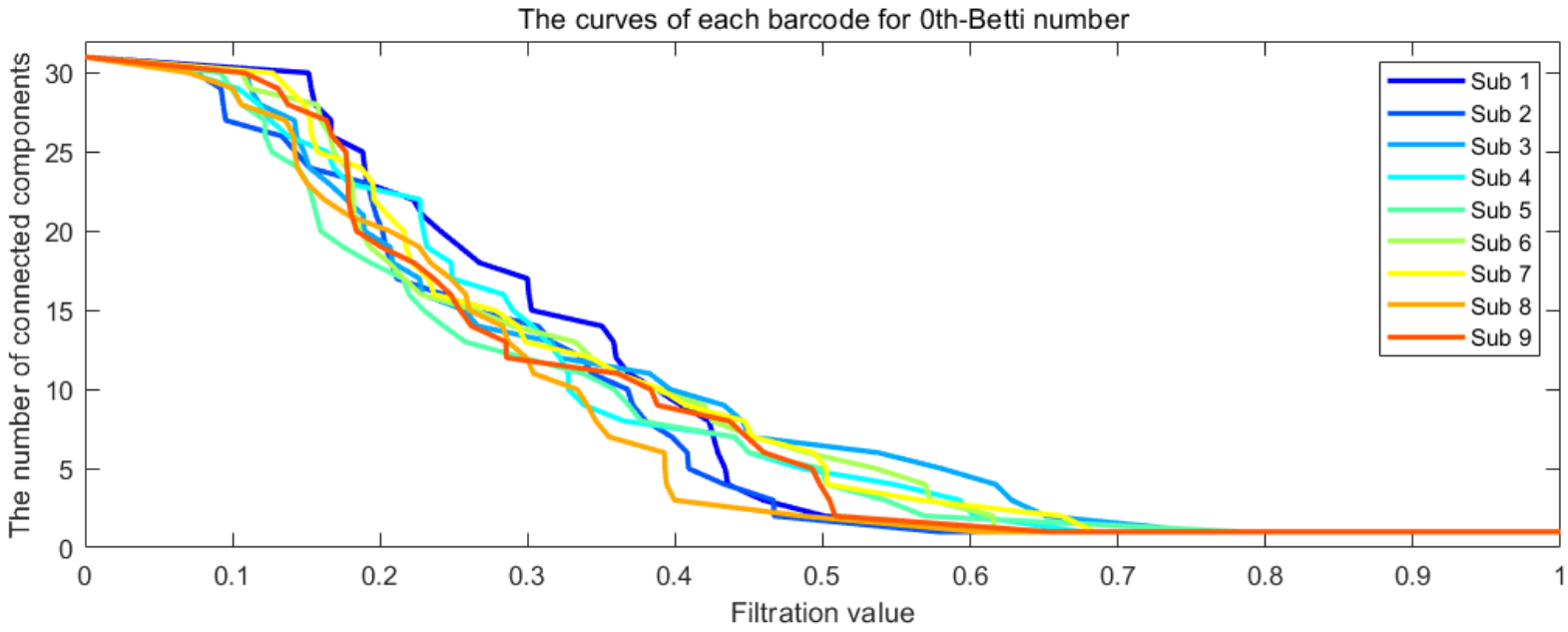
4.2 Result

- **We use the correlation-based distance as a metric**:
  - $\boldsymbol{d_C(x, y) = \left(1 - \left|\rho_{xy}\right|\right)^{1/2}}$ where $\boldsymbol{\rho_{xy}}$ is the Pearson correlation coefficient between $x \in X$ and $y \in Y$.



4.2 Result

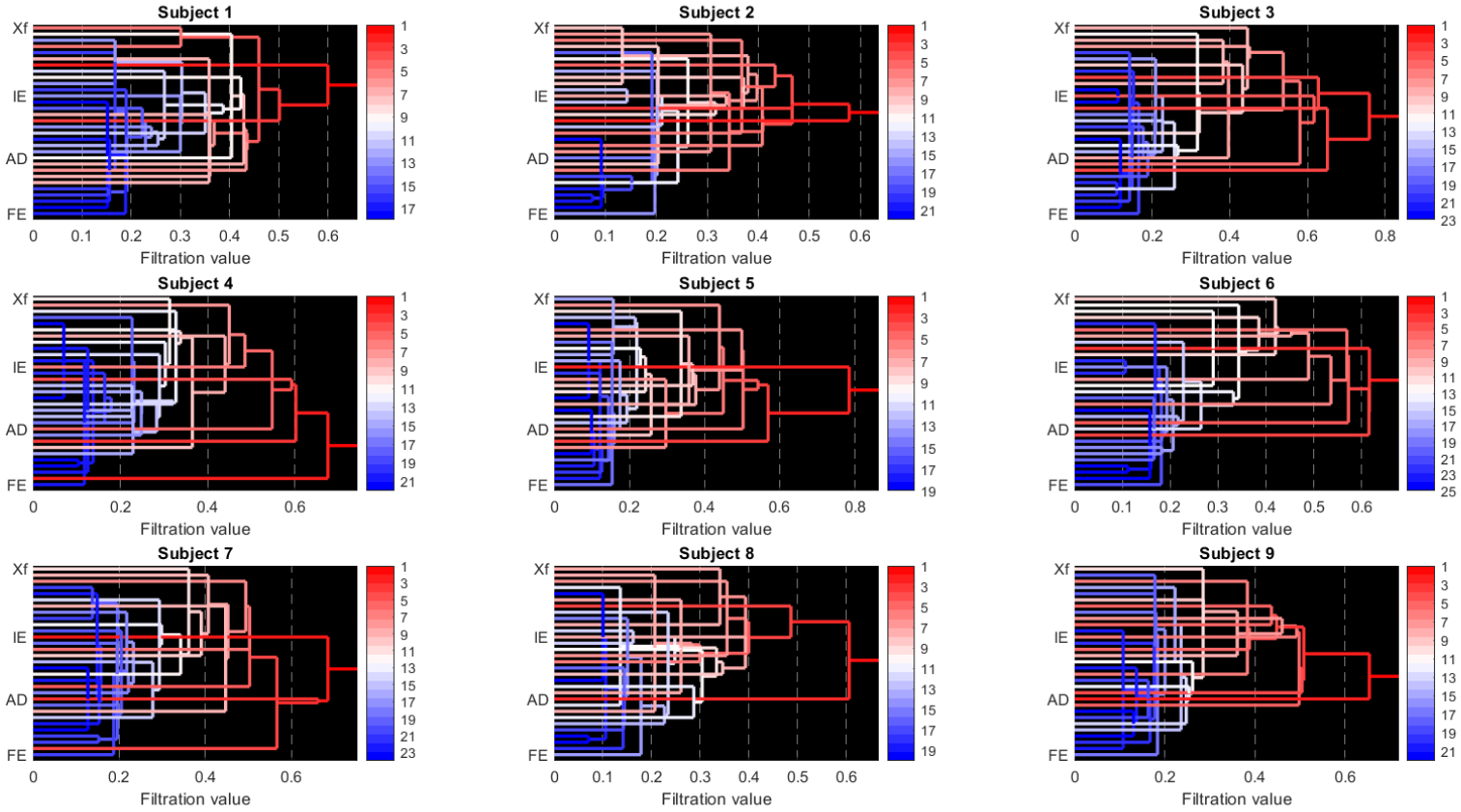
- **Barcode for β_0 as a curve representation**:
  - Subject 3 is the least connected and subject 8 is the fastest.
  - Only **1 through 10 connected components** (in 31 connected components) are **dominant at the difference between subjects**.



4.2 Result

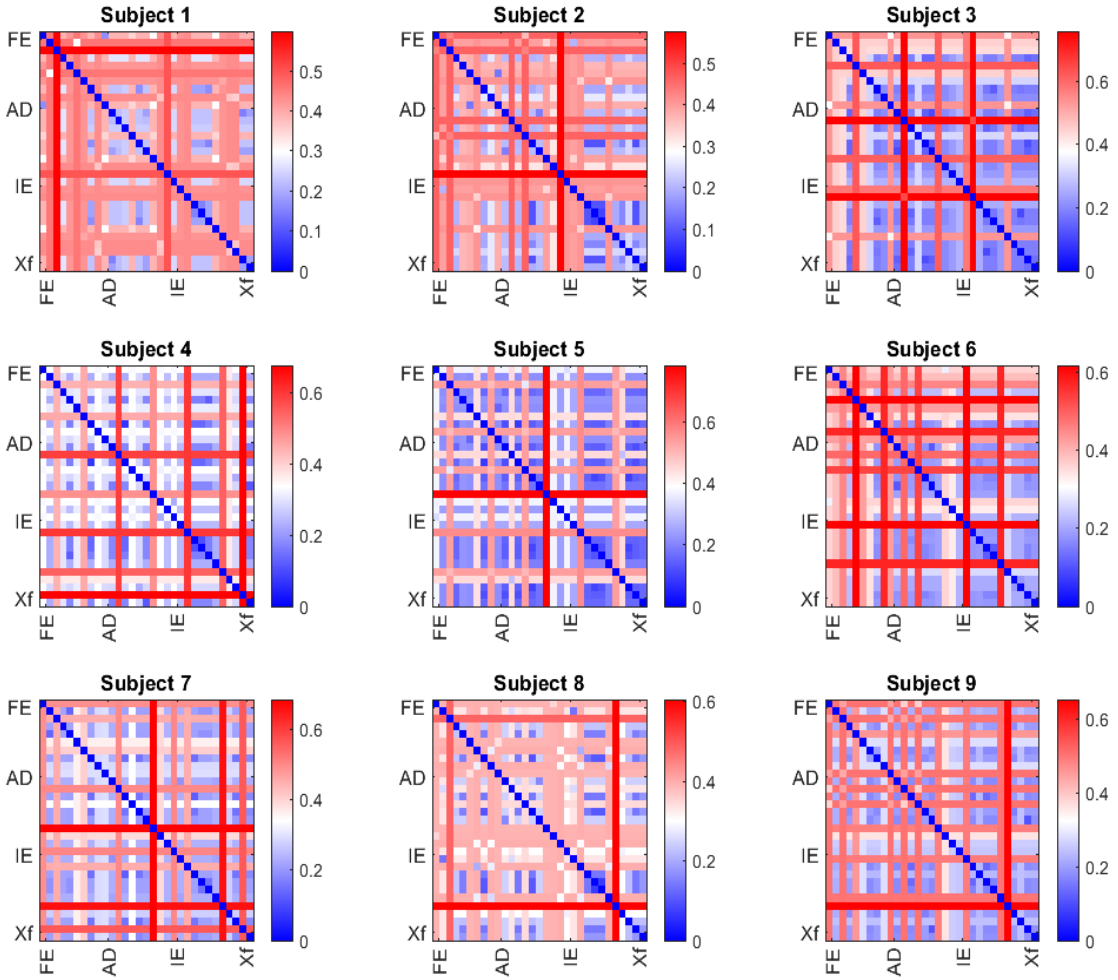
- **Single linkage dendrogram (SLD)**:

The **right body parts** (8,9,10,18,25,26,29,30) **cause the difference** between subjects.



4.2 Result

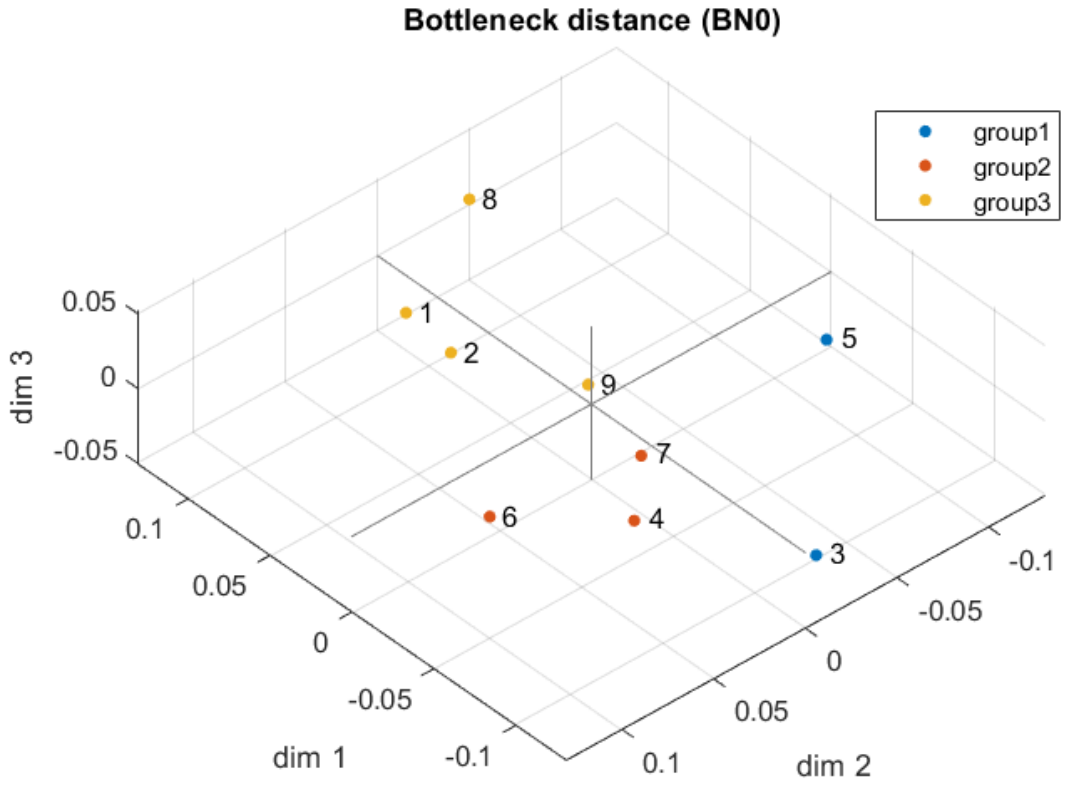
- **Single linkage matrix:**



4.2 Result

- **(MDS) Bottleneck distance for β_0** (goodness of fit = 0.072):

- Subject 4 and 7 are grouped and subject 1 and 2 are grouped

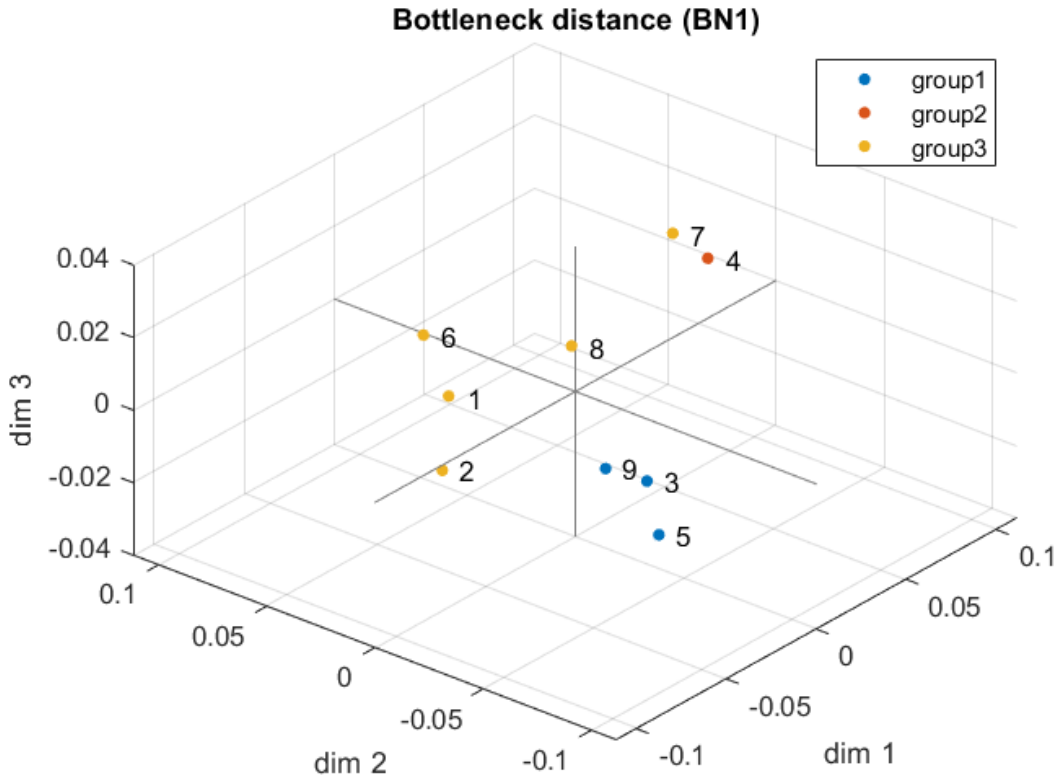


Clustering method : K-means clustering ($k = 3$)

4.2 Result

- **(MDS) Bottleneck distance for β_1** (goodness of fit = 0.0757):

- Subject 1 and 2 are grouped, and subject 4 and 7 are similar positions for each other.

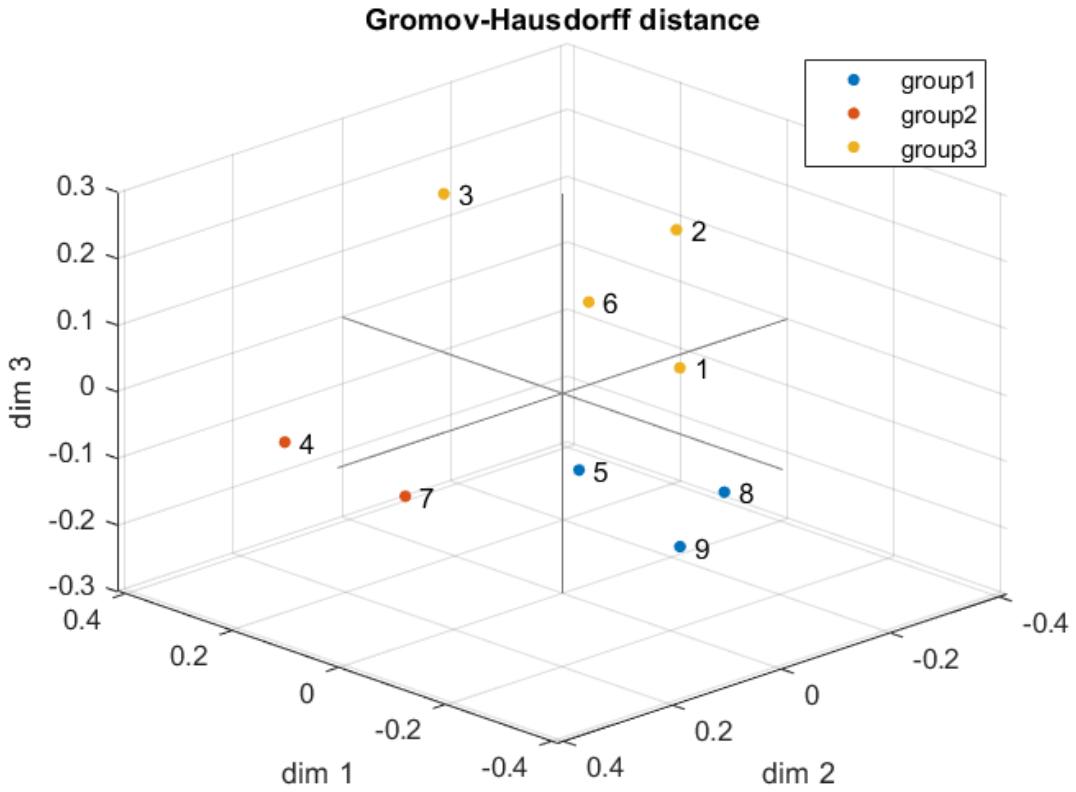


Clustering method : K-means clustering ($k = 3$)

4.2 Result

- **(MDS) Gromov-Hausdorff distance** (goodness of fit = 0.0963):

- Subject 4 and 7 are grouped and subject 1 and 2 are grouped



Clustering method : K-means clustering ($k = 3$)

Chapter 5. Conclusion

Chapter 5. Conclusion

✓ Contribution

- We applied TDA which is the new framework of data analysis based on the computational topology to the actual dataset. (with the conventional analysis methods)

- We developed the GUI program for TDA in the MATLAB environment. This program focused on the convenience to use and the accessibility

✓ Future work

- In this paper, we used only β_0 and β_1 as the topological invariant. (need to consider β_2 or other topological invariant.)

- We need to consider the other point cloud construction methods such as witness complex, alpha complex.

- The statistical analysis for the MDS result of each joint angle in the point clouds is also needed.

References

[1] Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

[2] Alpaydin, E. (2020). *Introduction to machine learning*. Massachusetts, USA: MIT press.

[3] Bengio, Y., Goodfellow, I., & Courville, A. (2017). *Deep learning* (Vol. 1). Massachusetts, USA: MIT press.

[4] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[5] Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society, 46*(2), 255-308.

[6] Mémoli, F. (2011, August). Metric structures on datasets: stability and classification of algorithms. In *International Conference on Computer Analysis of Images and Patterns* (pp. 1-33). Springer, Berlin, Heidelberg.

[7] Zerzucha, P., & Walczak, B. (2012). Concept of (dis) similarity in data analysis. *TrAC Trends in Analytical Chemistry, 38*, 116-128.

[8] Ghrist, R. (2008). Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society, 45*(1), 61-75.

[9] Edelsbrunner, H., & Harer, J. (2008). Persistent homology-a survey. *Contemporary mathematics, 453*, 257-282.

[10] Cohen-Steiner, D., Edelsbrunner, H., & Harer, J. (2007). Stability of persistence diagrams. *Discrete & computational geometry, 37*(1), 103-120.

[11] Adams, H., & Tausz, A. (2015). Javaplex tutorial. Retrieved from http://goo.gl/5uaRoQ

[12] Maria, C., Boissonnat, J. D., Glisse, M., & Yvinec, M. (2014, August). The Gudhi library: Simplicial complexes and persistent homology. In *International Congress on Mathematical Software* (pp. 167-174). Springer, Berlin, Heidelberg.

[13] Chazal, F., & Michel, B. (2017). An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *arXiv preprint arXiv:1710.04019*.

[14] Fasy, B. T., Kim, J., Lecci, F., & Maria, C. (2014). Introduction to the R package TDA. *arXiv preprint arXiv:1411.1830*.

References

[15] Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., & Harrington, H. A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science*, *6*(1), 17.

[16] Ferri, M. (2017). Persistent topology for natural data analysis—A survey. In *Towards Integrative Machine Learning and Knowledge Extraction* (pp. 117-133). Springer, Cham.

[17] Pun, C. S., Xia, K., & Lee, S. X. (2018). Persistent-Homology-based Machine Learning and its Applications--A Survey. *arXiv preprint arXiv:1811.00252*.

[18] Lee, H., Chung, M. K., Kang, H., Kim, B. N., & Lee, D. S. (2011, March). Discriminative persistent homology of brain networks. In *2011 IEEE international symposium on biomedical imaging: from nano to macro* (pp. 841-844). IEEE.

[19] Lee, H., Kang, H., Chung, M. K., Kim, B. N., & Lee, D. S. (2012). Persistent brain network homology from the perspective of dendrogram. *IEEE transactions on medical imaging*, *31*(12), 2267-2277.

[20] Carlsson, G., & Mémoli, F. (2010). Characterization, stability and convergence of hierarchical clustering methods. *The Journal of Machine Learning Research*, *11*, 1425-1470.

[21] Mémoli, F. (2008, June). Gromov-Hausdorff distances in Euclidean spaces. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1-8). IEEE.

[22] Borg, I., Groenen, P. J., & Mair, P. (2012). *Applied multidimensional scaling*. Springer Science & Business Media.

[23] Cho, K. D., Lee, E. J., Seo, T. H., Kim, K. R., & Koo, J. Y. (2012). General Research; Visualization of Bottleneck Distances for Persistence Diagram. *응용통계연구*, *25*(6), 1009-1018.

[24] Gamble, J., & Heo, G. (2010). Exploring uses of persistent homology for statistical analysis of landmark-based shape data. *Journal of Multivariate Analysis*, *101*(9), 2184-2199.

References

[25] Delory, B. M., Li, M., Topp, C. N., & Lobet, G. (2018). archiDART v3. 0: A new data analysis pipeline allowing the topological analysis of plant root systems. *F1000Research, 7*.

[26] Costa, J. P., & Škraba, P. (2015). A topological data analysis approach to the epidemiology of influenza. In *SIKDD15 Conference Proceedings*.

[27] Hajij, M., Jonoska, N., Kukushkin, D., & Saito, M. (2018). Graph based analysis for gene segment organization in a scrambled genome. *arXiv preprint arXiv:1801.05922*.

[28] Zomorodian, A. (2005). *Topology for computing* (Vol. 16). Cambridge university press.

[29] Zomorodian, A. (2010). Fast construction of the Vietoris-Rips complex. *Computers & Graphics, 34*(3), 263-271.

[30] Edelsbrunner, H., & Harer, J. (2010). *Computational topology: an introduction.* American Mathematical Soc.

[31] De Silva, V., & Ghrist, R. (2007). Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology, 7*(1), 339-358.

[32] Hatcher, A. (2002). *Algebraic topology.* Cambridge university press.

[33] Aktas, M. E., Akbas, E., & El Fatmaoui, A. (2019). Persistence homology of networks: methods and applications. *Applied Network Science, 4*(1), 61.

[34] Moon, C., Giansiracusa, N., & Lazar, N. A. (2018). Persistence terrace for topological inference of point cloud data. *Journal of Computational and Graphical Statistics, 27*(3), 576-586.

[35] Zomorodian, A., & Carlsson, G. (2005). Computing persistent homology. *Discrete & Computational Geometry, 33*(2), 249-274.

[36] Chung, M. K., Lee, H., DiChristofano, A., Ombao, H., & Solo, V. (2019). Exact topological inference of the resting-state brain networks in twins. *Network Neuroscience, 3*(3), 674-694.

References

[37] Lee, H., Chung, M. K., Kang, H., Kim, B. N., & Lee, D. S. (2011, September). Computing the shape of brain networks using graph filtration and Gromov-Hausdorff metric. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 302-309). Springer, Berlin, Heidelberg.

[38] Wickelmaier, F. (2003). An introduction to MDS. *Sound Quality Research Unit, Aalborg University, Denmark*, *46*(5), 1-26.

[39] Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, *100*(5), 401-409.

[40] Chen, J., Ng, Y. K., Lin, L., Jiang, Y., & Li, S. (2019). On triangular Inequalities of correlation-based distances for gene expression profiles. *bioRxiv*, 582106.