The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.6.1009
25(6), 1009–1018

Visualization of Bottleneck Distances for Persistence Diagram

Kyu-Dong Cho[1] · Eunjee Lee[2] · Taehee Seo[3] · Kwang-Rae Kim[4] · Ja-Yong Koo[5]

[1]Department of Statistics, Korea University
[2]Statistics and Operations Research, University of North Carolina at Chapel Hill
[3]Department of Statistics, Korea University
[4]Institute for Mathematical Stochastics, Georg-August-University of Goettingen
[5]Department of Statistics, Korea University

(Received August 30, 2012; Revised September 25, 2012; Accepted October 31, 2012)

Abstract

Persistence homology (a type of methodology in computational algebraic topology) can be used to capture the topological characteristics of functional data. To visualize the characteristics, a persistence diagram is adopted by plotting baseline and the pairs that consist of local minimum and local maximum. We use the bottleneck distance to measure the topological distance between two different functions; in addition, this distance can be applied to multidimensional scaling(MDS) that visualizes the imaginary position based on the distance between functions. In this study, we use handwriting data (which has functional forms) to get persistence diagram and check differences between the observations by using bottleneck distance and the MDS.

Keywords: Multidimensional scaling(MDS), persistent homology, persistence diagram, bottleneck distance, functional data, topology.

1. 서론

여러 실험으로부터 얻어지는 자료의 분석방법으로 전통적인 일변량/다변량 분석 방법이 활발히 진행되어 왔다. 하지만 실험 구조의 복잡성이나 그 측정 자료가 지닌 내재적 의미를 깊게 파악하기 위하여 새로운 접근이나 분석 방법이 대두되고 있다. 특히 대수적(algebraic) 위상 논법에 기반을 둔 계산 위상(computational topology)의 관점에서 자료를 분석함으로써 표면적뿐만 아닌 잠재적인 구조를 동시에 파악함으로써 전체적인 현상을 설명할 수 있다.

지속 호몰로지(persistent homology)는 전산 대수 위상(computational algebraic topology)의 한 방법론으로써, 함수의 위상적 특성을 측정한다는 기본적인 관점을 출발점으로 하고 있다. 특히 자료가 부드러운 함수의 형태로 주어진 경우에 두 가지 주된 관심 사항이 있는데, 위상적 특성이 발생(appearance,

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2011-0002450).
[5]Corresponding author: Professor, Department of Statistics, Korea University, 5-1 Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea. E-mail: jykoo@korea.ac.kr

birth)하는 시점의 함수 값과 소멸(disappearance, death)하는 시점의 함수 값이다. 이 두 함수 값을 짝지어 함수의 위상적 특성을 도식화하여 나타낸 도표를 지속 다이어그램(persistence diagram)이라고 한다 (Edelsbrunner와 Harer, 2008). 지속 다이어그램은 함수 형태의 자료가 갖는 위상적 특성이라는 수학적인 개념을 밝히는데 있어서 중요한 도구로 이용되어 왔다.

Edelsbrunner와 Harer (2008)는 지속 호몰로지의 등장 배경과 분석 방법에 대하여 설명하고 있다. 국소 임계값들을 짝지어 분석하는 연구로는 Edelsbrunner와 Harer (2008), Edelsbrunner 등 (2002), Zomorodian와 Carlsson (2005) 등이 있다. 유한 개의 국소 임계값들을 비선형 모형을 통하여 짝을 지어 도시하는 지속 다이어그램에 대한 연구로는 Cohen-Steiner 등 (2007), Edelsbrunner와 Harer (2008), Morozov (2008), Zomorodian (2001) 등이 있다. 최근에는 Bubenik과 Kim (2007), Gamble과 Heo (2010), Chung 등 (2009)처럼 통계학 관점에서 지속 호몰로지가 연구되고 있다.

다차원 척도법(MDS; multidimensional scaling)은 다수의 개체 간 유사성이나 물리적 거리 등의 비유사성을 저차원으로 표현해내는 통계적 기법으로 Borg와 Groenen (2005)에서 개념과 통계적 성질을 자세히 다루고 있다. MDS는 거리로 표현된 유사성과 비유사성을 시각적으로 볼 수 있다는 장점이 있으며 마케팅이나 통신, 네트워크분석 등 사회 전반적인 분야에서 많이 사용되고 있다. 자료의 형태가 밀도함수인 경우 Delicado (2011a)와 Delicado (2011b)에서 MDS를 사용하였으며, 특히 Delicado (2011b)에서는 함수형자료에서 MDS를 이용하면 주성분분석(principal component analysis) 보다 자연스러운 차원축소의 결과를 얻는다고 하였다.

본 논문에서는 지속 다이어그램이 갖는 특성에 초점을 맞추어 병목거리를 다차원 척도법으로 시각화하고자 한다. 2장에서는 지속 다이어그램의 정의와 해석 방법을 소개하고, 병목 거리를 이용한 다차원 척도법을 소개하며, 3장에서는 모의실험을 통해 유클리드 거리에 비해 병목 거리가 갖는 이점을 알아보고, 이를 바탕으로 함수의 위상적인 특징을 살펴보기 위해 손글씨(handwriting) 자료를 분석하였다. 마지막으로 4장에서는 본 연구의 결과와 향후 과제에 대해 논의하고자 한다.

2. 지속 다이어그램과 병목 거리를 활용한 다차원 척도법

2.1. 지속 다이어그램의 정의와 해석

함수 f를 유한개의 임계값을 갖는 모스 함수(Morse function)라고 하고, f의 국소 최소값(local minimum)들을 $x_1, x_2, \ldots, x_n$, 국소 최대값(local maximum)들을 $y_1, y_2, \ldots, y_m$라고 하자. 이때 모스 함수의 특성으로부터 다음과 같이 정렬 가능하다.

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)},\ y_{(1)} < y_{(2)} < \cdots < y_{(m)}.$$

또한 $x_1, x_2, \ldots, x_n$과 $y_1, y_2, \ldots, y_m$을 호몰로지컬(homological) 임계값이라고 정의한다. 특정 임계값, c에 대해서 다음과 같은 서브레벨(sublevel) 집합 $S(c)$를 정의할 수 있다.

$$S(c) = f^{-1}(-\infty, c].$$

이 때 $S(c)$는 겹치지 않는 집합들로 구성되는데, 이러한 집합을 요소(component)라고 하고 요소들의 갯수를 $\#S(c)$로 표기하자. 그러면 특정 최소값 x와 충분히 작은 실수 ε에 대해

$$\#S(x) = \#S(x - \varepsilon) + 1$$

로 새로운 요소가 발생(birth)하고, 특정 최대값 y에 대해

$$\#S(y) = \#S(y - \varepsilon) + 1$$

로 요소가 소멸(death)한다.

만약 어떤 요소가 최소값 x_i에서 발생하여 최대값 y_j에서 소멸하면, (x_i, y_j)를 쌍(pair)이라고 하고, $y_j - x_i$를 이 요소의 지속(persistence)이라고 한다 (Edelsbrunner와 Harer, 2008). 함수 f에 대해 이러한 가능한 모든 쌍을 $\bar{\mathcal{D}}(f)$라 표기하며, 이를 2차원 상에 도표화 한 것을 지속 다이어그램(persistence diagram)이라 한다 (Edelsbrunner와 Harer, 2008; Chung 등, 2009).

모든 실수점에서 하나의 요소가 발생하는 동시에 소멸한다고 가정하자. 임의의 실수점을 a라고 하면, 이 요소에 대한 쌍은 (a, a)가 될 것이다. 이런 쌍을 모두 모아놓은 집합이 원점을 지나는 기울기가 1인 직선이며 이 때의 요소들의 지속은 모두 0이다. 그러므로 원점을 지나는 기울기가 1인 직선을 기준선(baseline)으로 함께 도시하며 지속 호몰로지에 기준선이 포함된 집합을 $\mathcal{D}(f)$로 표기한다. 기준선에서 먼 쌍일수록 그 요소의 지속이 길다고 할 수 있고, 기준선에 가까운 쌍일수록 지속이 짧은 요소라고 할 수 있다. 여기에서 요소의 지속이 짧다는 것은 다시 데이터를 관측했을 때 그 요소가 사라질 가능성이 높다는 것을 의미한다. 따라서 기준선에 가까운 쌍은 함수의 거짓 파동일 가능성이 크고, 반대로 기준선과 먼 쌍일수록 유의미한 파동을 나타낸다고 볼 수 있다. 또한 지속 다이어그램의 쌍의 갯수는 데이터가 얼마나 많은 변동을 갖는지 나타내는 척도로 기준선 가까이에 있는 점들을 제외한 다른 점들의 갯수가 함수의 유의미한 변동을 수량화한다고 볼 수 있다.

2.2. 병목 거리를 이용한 다차원 척도법

$\mathbb{R} \to \mathbb{R}$에서 정의된 함수 f와 g를 고려하자. 그러면 f와 g의 지속 다이어그램 사이의 거리는 병목 거리(bottleneck distance)로 정의되며 $\gamma : \mathcal{D}(f) \to \mathcal{D}(g)$일 때

$$d(\mathcal{D}(f), \mathcal{D}(g)) = \inf_{\gamma} \sup_{p \in \mathcal{D}(f)} \|p - \gamma(p)\|_\infty$$

와 같다 (Chung 등, 2009).

2 차원상에서 정의된 병목 거리는 다음과 같이 나타낼 수 있다. 먼저 $(x_1^f, y_1^f), \ldots, (x_m^f, y_m^f)$와 (x_1^g, y_1^g), $\ldots, (x_n^g, y_n^g)$를 각각 $\bar{\mathcal{D}}(f)$와 $\bar{\mathcal{D}}(g)$의 값이라 하자.

우선 $m = n$을 가정하고, P_m을 $\{1, \ldots, m\}$의 순열 집합(set of permutation)이라 할 때 $\pi \in P_m$을 고려하고 다음을 정의하자.

$$B(i, \pi) = \max\left\{\left|x_i^f - x_{\pi(i)}^g\right|, \left|y_i^f - y_{\pi(i)}^g\right|\right\}.$$

그러면 $B(\pi)$는

$$B(\pi) = \max_{i=1,\ldots,m} B(i, \pi)$$

와 같이 최대값으로 나타내며, 병목 거리 B는

$$B = \min_{\pi \in P_m} B(\pi) \tag{2.1}$$

로 구할 수 있다.

다음으로 $m \neq n$이면, $m = n$인 경우와 비슷하게 구할 수 있다. 먼저 $m > n$을 가정하고 $\pi \in P_m$를 고려하자. $m < n$이면 f와 g를 바꿀 수 있다. 그러면,

$$\left(x_{\pi(i)}^g, y_{\pi(i)}^g\right) = \begin{cases} \left(x_{\pi(i)}^g, y_{\pi(i)}^g\right), & \text{if } \pi(i) = 1, \ldots, n, \\ \left(\alpha_{\pi(i)}, \alpha_{\pi(i)}\right), & \text{if } \pi(i) = n+1, \ldots, m \end{cases}$$

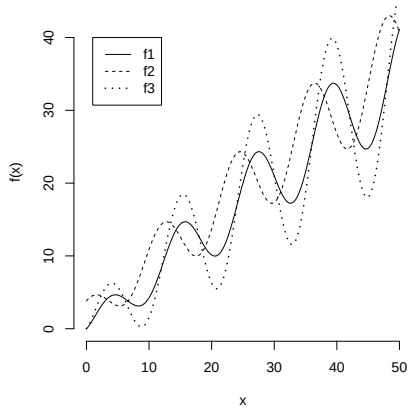


Figure 3.1. Simulated data

이다. 여기서 $\alpha_{\pi(i)} = (x_i^f + y_i^f)/2$는 f에서 지속 다이어그램의 값에 대응하는 가장 가까운 값이며, (x_i^f, y_i^f)에서 기준선에 직교 사영(orthogonal projection)을 한 값이다. 따라서 $\pi(i) = 1, \ldots, m$에 대해

$$B(i,\pi) = \max\left\{\left|x_i^f - x_{\pi(i)}^g\right|, \left|y_i^f - y_{\pi(i)}^g\right|\right\}$$

를 구할 수 있고

$$B(\pi) = \max_{i=1,\ldots,m} B(i,\pi)$$

에 따라 병목 거리 B가

$$B = \min_{\pi \in P_m} B(\pi) \tag{2.2}$$

와 같이 정의된다.

다차원 척도법이란 대상간의 유사성(또는 선호도) 측도에 의거해서 대상을 저차원 공간 속에 배치시키는 방법으로, 여기에서 병목 거리를 이용하면 두 함수형 자료의 위상학적 거리를 시각적으로 표현할 수 있다. 다차원 척도법은 데이터 속에 잠재해있는 패턴, 구조를 찾아낼 수 있을 뿐 아니라 그 구조를 저차원의 공간에 기하학적으로 표현하기 때문에 병목 거리와 다차원 척도법을 통해 함수형 자료의 유사성을 볼 수 있다.

3. 자료 분석

3.1. 모의 실험

병목 거리가 유클리드 거리에 비해 갖는 이점을 보기 위해 다음과 같은 3가지의 함수형 자료를 생성하였다.

$$f_1(x) = 0.7x + x^{\frac{1}{2}} \sin\left(\frac{1}{6}\pi x\right) + \varepsilon$$

$$f_2(x) = 0.7(x+3) + (x+3)^{\frac{1}{2}} \sin\left(\frac{1}{6}\pi(x+3)\right) + \varepsilon$$

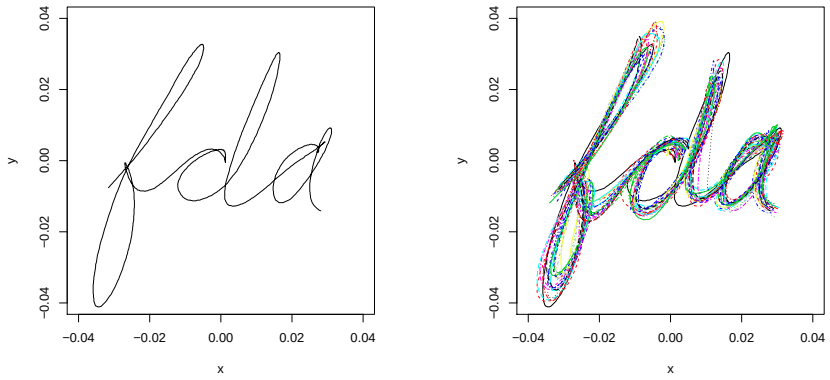
$$f_3(x) = 0.7x + 2x^{\frac{1}{2}} \sin\left(\frac{1}{6}\pi x\right) + \varepsilon$$

Table 3.1. The Uclidean distances matrix for functions f_1, f_2, f_3

	f_2	f_3
f_1	178.237	110.096
f_2	0	260.73

Table 3.2. The bottleneck distances matrix for functions f_1, f_2, f_3

	f_2	f_3
f_1	0.009	6.624
f_2	0	6.622



(a) Observation no.1 (b) Observation no.1∼20

Figure 3.2. Raw data

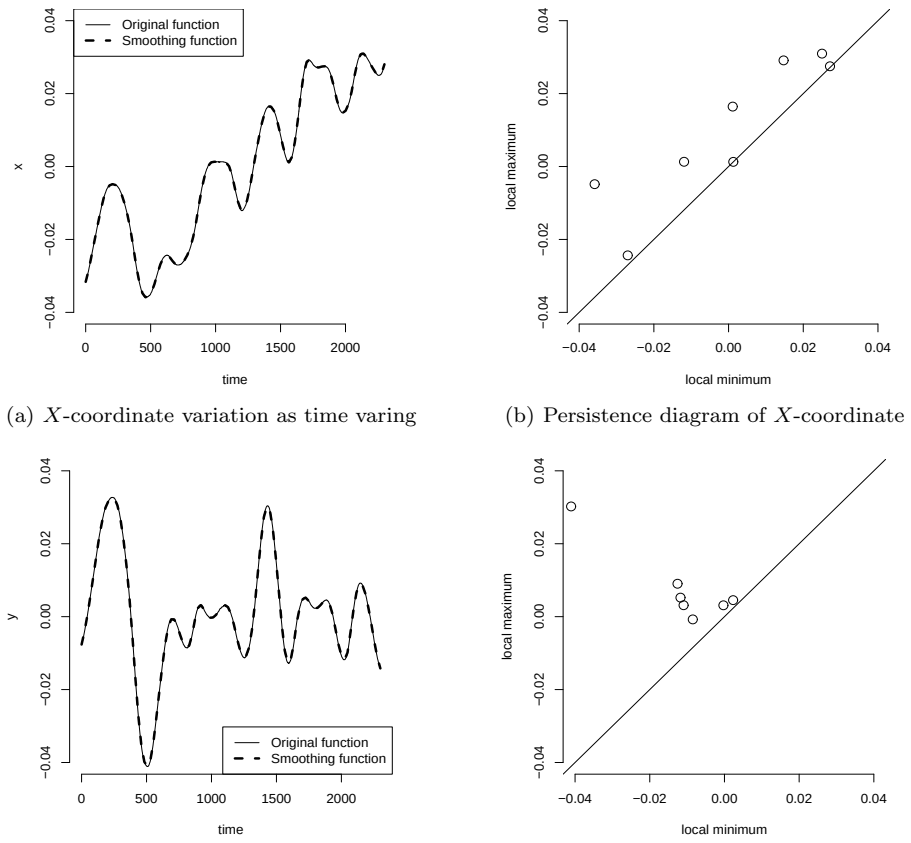
여기서 $\varepsilon \sim \text{Normal}(0, 0.01)$이다. Figure 3.1에서 보여주듯이 f_2는 f_1을 3만큼 왼쪽으로 대칭이동한 것이고 f_3은 f_1의 진동폭을 2배 만큼 증가시킨 함수이다. f_1과 f_2는 함수형 자료를 위상학적 관점에서 동일한 함수로 볼 수 있으며, f_3은 f_1과 f_2에 비해 폭이 더 크므로, 호몰로지 관점에서는 다른 형태를 지닌 함수로 볼 수 있다.

Table 3.1은 함수 간 각 점들을 유클리드 거리로 나타낸 것이다. 위상적으로 동일한 f_1과 f_2의 거리보다 f_2와 f_3의 거리가 더 가깝게 나타나는 것을 볼 수 있다. 이것은, 유클리드 거리와 같은 일반적인 거리로는 함수 형태의 위상학적 차이를 보기는 어렵다는 것을 알 수 있다. 반면, Table 3.2의 병목 거리 행렬을 보면, 함수 f_1과 f_2의 거리가 거의 0에 가까우며, 함수 f_2, f_3간의 거리와 함수 f_1, f_3간의 거리가 거의 동일함을 알 수 있다. 즉, 병목거리가 두 함수의 위상적 차이을 잘 설명해주고 있음을 보여준다. 따라서 위상적 특성이 중요한 함수형 자료를 분석할 때, 병목거리를 도입하는 것이 보다 적절하다.

3.2. 손글씨 자료

본 논문에서 사용된 자료는 Ramsay와 Silverman (2005)에 있는 손글씨 데이터로, 'fda'라는 글자를 필기체로 20번 작성한 데이터이다. 각 개체는 전체 2.3초 동안 1.642857ms 등간격으로 1401번 측정되었으며, 매 단위마다 X, Y좌표를 기록한 것이다. 자료는 R 소프트웨어의 'fda' 패키지에서 구할 수 있으며, 자료에 대한 자세한 설명도 볼 수 있다.

Figure 3.2의 (a)는 1번 시행에 대한 그림이며, (b)는 전체 자료에 대한 그림이다. Figure 3.3의 (a)와



(a) X-coordinate variation as time varing

(b) Persistence diagram of X-coordinate

(c) Y-coordinates variation as time varing

(d) Persistence diagram of Y-coordinate

Figure 3.3. Persistence diagram

(c)는 각각 1번 자료에 대해 시간에 따른 X좌표와 Y좌표에 대한 그림이다. 필기체 인식과정에서 발생하는 측정오차를 보정하기 위해, R의 'LOWESS' 함수를 이용하여 자료를 평활(smoothing)하였으며, 실선이 원자료, 점선이 함수를 평활한 자료이다. 평활과정에서 자료의 형태를 유지하기 위하여 평활 모수를 0.01을 사용하였다. (b)와 (d)는 각 축 좌표에 대한 지속 다이어그램이다. 8개의 지속이 존재하는 X축이 7개의 지속이 존재하는 Y축보다 파동이 많다. 여기서 기준선과 가까운 X축 3개와 Y축 2개의 지속은 거짓파동일 가능성이 높다. X축과 Y축간의 식 (2.1)과 (2.2)에 의한 병목 거리는 0.0291이다.

Table 3.3과 Table 3.4는 20개 자료에 대해서 X축과 Y축 별로 지속 쌍들의 병목 거리를 나타낸다. X축에서 가장 먼 병목거리를 갖는 두 개체는 7번과 17번으로 이들 사이의 병목 거리는 0.0154이다. 마찬가지로 Y축에 대해서 가장 먼 개체는 1번과 7번 또는 1번과 13번으로 이들 사이의 병목 거리는 0.0099이다. 거리의 개념은 상대적이어야 하는데 이는 다른 개체와의 거리도 함께 고려되어야 한다. 하지만 거리 행렬 차원이 크기 때문에 차원 축소의 개념 중 병목 거리에 대한 시각화 방법인 MDS를 수행하였다. Figure 3.4는 X축 Y축 각각 병목 거리에 대한 MDS를 나타내며, Figure 3.4에서 괄호부분은 stress를 나타낸다. X축에서 7, 9, 12, 17, 18번 개체가 다른 개체에 비해 위상적 거리가 먼 것을 알 수 있으며, Y축에서는 병목 거리 행렬과 마찬가지로 1번 개체와 7, 13번 개체가 위상적 거리가 많이 떨어져 있다.

Table 3.3. The bottleneck distances for X-coordinate functional data

	2	3	4	5	6	7	8	9	10	11
1	0.0038	0.0037	0.0043	0.0048	0.0066	0.0051	0.0033	0.0094	0.0056	0.0051
2		0.0048	0.0043	0.0050	0.0037	0.0066	0.0037	0.0056	0.0049	0.0051
3			0.0038	0.0070	0.0046	0.0028	0.0032	0.0057	0.0049	0.0071
4				0.0069	0.0045	0.0038	0.0038	0.0063	0.0048	0.0069
5					0.0023	0.0068	0.0041	0.0065	0.0050	0.0048
6						0.0058	0.0034	0.0042	0.0040	0.0039
7							0.0050	0.0064	0.0048	0.0069
8								0.0062	0.0032	0.0042
9									0.0050	0.0066
10										0.0021

	12	13	14	15	16	17	18	19	20
1	0.0042	0.0046	0.005	0.0053	0.0039	0.0052	0.0047	0.0041	0.0041
2	0.0052	0.0044	0.0037	0.0055	0.0054	0.0046	0.0037	0.0036	0.0046
3	0.0038	0.0050	0.0046	0.0044	0.0048	0.0066	0.0032	0.0046	0.0059
4	0.0037	0.0026	0.0045	0.0043	0.0047	0.0065	0.0031	0.0045	0.0058
5	0.0032	0.0056	0.0024	0.0056	0.0055	0.0032	0.0048	0.0037	0.0047
6	0.0025	0.0043	0.0031	0.0046	0.0046	0.0020	0.0024	0.0042	0.0038
7	0.0146	0.0058	0.0048	0.0043	0.0137	0.0154	0.0049	0.0045	0.0057
8	0.0016	0.0042	0.0030	0.0038	0.0038	0.0037	0.0036	0.0038	0.0032
9	0.0144	0.0048	0.0043	0.0041	0.0122	0.0148	0.0047	0.0063	0.0054
10	0.0026	0.0036	0.0030	0.0032	0.0043	0.0038	0.0035	0.0020	0.0032
11	0.0032	0.0057	0.0029	0.0027	0.0034	0.0032	0.0048	0.0029	0.0027
12		0.0025	0.0017	0.0025	0.0025	0.0028	0.0145	0.0025	0.0022
13			0.0034	0.0030	0.0036	0.0052	0.0025	0.0033	0.0045
14				0.0036	0.0036	0.0032	0.0035	0.0023	0.0031
15					0.0030	0.0028	0.0039	0.0030	0.0022
16						0.0033	0.0123	0.0035	0.0014
17							0.0149	0.0034	0.0025
18								0.0034	0.0039
19									0.0025

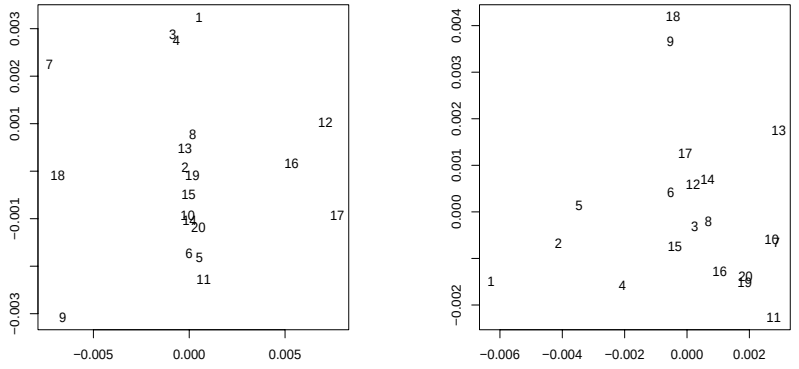
X축 및 Y축 병목 거리를 동시에 살펴보기 위한 방법으로 두 거리 행렬의 평균 및 최대, 최소값을 고려할 수 있다. 최대값을 고려하였을 경우 위상적으로 동질적인 개체를 판별할 때 용이하며, 반대로 최소값을 고려하면 위상적으로 이질적인 개체판별이 용이하다. 그리고 각 축의 정보를 모두 사용하는 평균값을 사용하면 개체들간의 종합적인 동질성을 판단할 수 있다. 본 논문에서는 두 거리의 평균으로 MDS 및 계층적 군집화(hierarchical clustering)을 수행하여 Figure 3.5에 나타내었다. 여기서 계층적 군집화 연결방법은 평균 연결법을 사용하였다. MDS와 계층적 군집의 나무형그림(dendrogram) 모두 1번 및 2번 개체와 9번 및 18번 개체가 서로는 가깝지만 다른 개체와는 병목 거리가 먼 것을 알 수 있다.

MDS에서 살펴보면 19번 및 20번 개체는 가깝고 7번 및 17번 개체는 먼 것을 알 수 있는데 이러한 이유를 개별적 자료를 통해 확인하고자 한다. Figure 3.6에서 이들 4개의 개체를 나타내었다. Figure 3.6의 (a)에서 보면 19번 개체와 20번 개체는 'f'와 'a'의 전체적인 모습이 일치하며, 'd'의 Y축 높이도 일치한다. 특히 X축으로의 퍼진 정도와 Y축으로의 높이가 거의 일치한다고 보여진다. 이는 Figure (b)에서 지속 다이어그램을 살펴봐도 각 점들이 매우 가깝게 위치한 것을 알 수 있다. 반면, (c)에서 7번을 살

Table 3.4. The bottleneck distances for Y-coordinate functional data

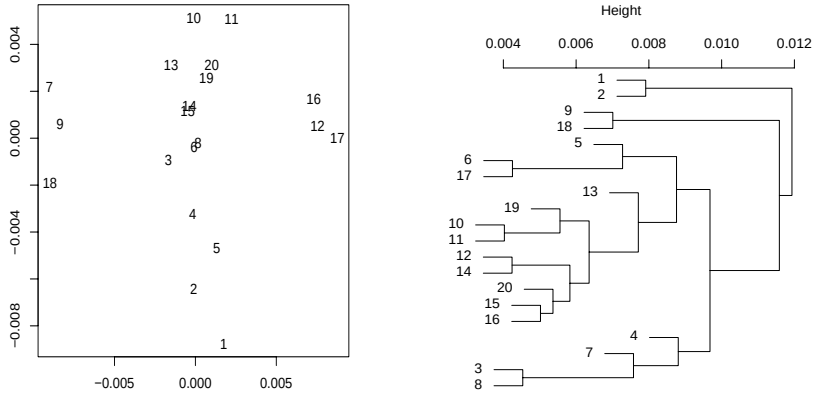
	2	3	4	5	6	7	8	9	10	11
1	0.0041	0.0073	0.0065	0.0044	0.0072	0.0099	0.0081	0.0081	0.0092	0.0096
2		0.0056	0.0043	0.0057	0.0054	0.0080	0.0064	0.0062	0.0073	0.0083
3			0.0035	0.0053	0.0035	0.0039	0.0013	0.0048	0.0043	0.0057
4				0.0036	0.0036	0.0075	0.0040	0.0056	0.0068	0.0050
5					0.0049	0.0074	0.0061	0.0056	0.0067	0.0076
6						0.0057	0.0031	0.0043	0.0050	0.0046
7							0.0034	0.0068	0.0043	0.0057
8								0.0056	0.0049	0.0063
9									0.0057	0.0071
10										0.0019

	12	13	14	15	16	17	18	19	20
1	0.0069	0.0099	0.0073	0.0059	0.0080	0.0078	0.0087	0.0088	0.0085
2	0.0050	0.0080	0.0062	0.0045	0.0063	0.0061	0.0068	0.0071	0.0068
3	0.0020	0.0039	0.0036	0.0020	0.0027	0.0043	0.0053	0.0041	0.0030
4	0.0044	0.0075	0.0049	0.0035	0.0032	0.0031	0.0062	0.0041	0.0044
5	0.0044	0.0074	0.0048	0.0040	0.0060	0.0041	0.0062	0.0068	0.0065
6	0.0031	0.0057	0.0031	0.0044	0.0039	0.0023	0.0045	0.0055	0.0036
7	0.0039	0.0047	0.0043	0.0040	0.0043	0.0056	0.0072	0.0051	0.0039
8	0.0027	0.0039	0.0042	0.0026	0.0034	0.0033	0.0060	0.0050	0.0035
9	0.0028	0.0037	0.0033	0.0035	0.0055	0.0029	0.0023	0.0063	0.0060
10	0.0030	0.0037	0.0024	0.0035	0.0036	0.0041	0.0061	0.0027	0.0024
11	0.0044	0.0051	0.0038	0.0037	0.0030	0.0048	0.0076	0.0036	0.0029
12		0.0030	0.0025	0.0018	0.0026	0.0038	0.0036	0.0034	0.0031
13			0.0040	0.0040	0.0043	0.0053	0.0051	0.0042	0.0039
14				0.0039	0.0034	0.0024	0.0038	0.0035	0.0027
15					0.0021	0.0052	0.0050	0.0037	0.0035
16						0.0047	0.0059	0.0041	0.0038
17							0.0035	0.0040	0.0037
18								0.0067	0.0064
19									0.0031



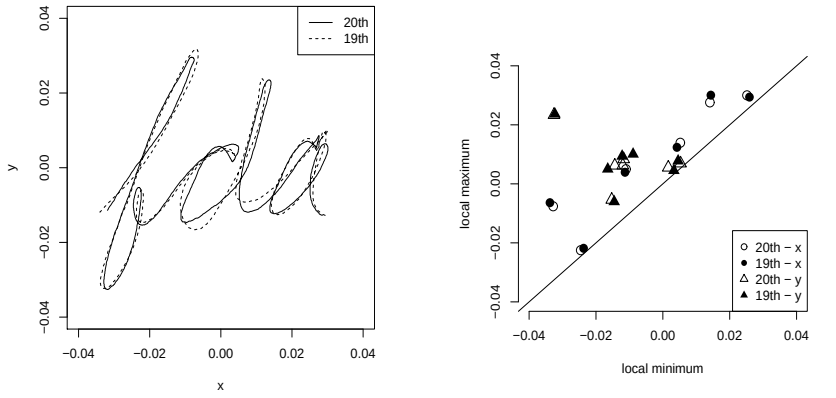
(a) X-coordinate functional data(0.260) (b) Y-coordinate functional data(0.087)

Figure 3.4. Multidimensional scaling

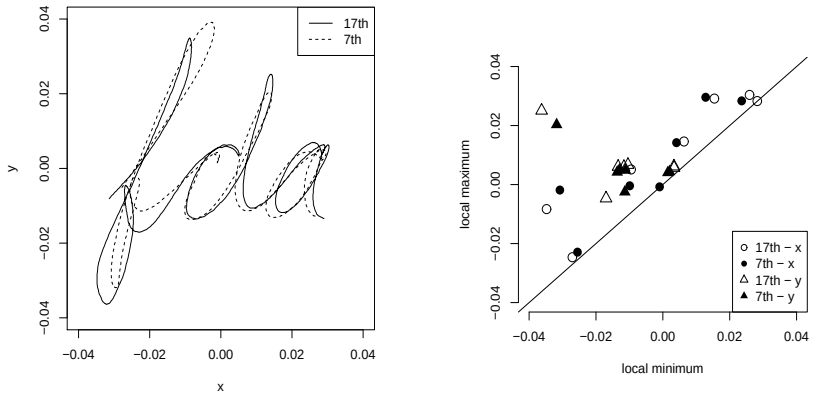


(a) Multidimensional scaling(0.117) (b) Hierarchical clustering

Figure 3.5. MDS and Hierarchical clustering for mean of two distance matrices



(a) Raw data (b) Persistence diagram for X and Y



(c) Raw data (d) Persistence diagram for X and Y

Figure 3.6. Comparison between observations

펴보면 'f'의 아랫부분이 X축 방향으로 좁게 나타나며, 윗부분은 넓적하고 'd'와의 연결부분은 Y축으로 높게 형성되어 있다. 그리고 'd'의 윗부분이 둥글며 'a'부분도 둥글게 쓰여있다. 그에 비해 17번 개체는 윗부분이 전체적으로 길죽하며, 'd'와 'a'역시 각 축별로 차이가 있다. (d)의 지속 다이어그램도 (b)보다는 점들의 위치가 멀다는 것을 확인할 수 있다.

4. 결론

본 연구에서는 함수형 자료의 위상적인 특성을 알아보기 위하여 지속 다이어그램 및 병목 거리를 이용하였으며, 이를 다차원 척도법에 적용하였다. 지속 다이어그램을 이용하여 시간에 따른 두 함수의 변동을 나타낸다면, 더이상 시간에 의존하지 않는 그래프를 얻을 수 있으며, 함수형 자료 분석에서 사용되는 정렬(registration)을 하지 않아도 위상적 특성을 살펴 볼 수 있는 장점이 있다. 또한 지속 호몰로지와 다차원 척도법의 접목은 함수형 자료의 거리가 가깝고 먼 정도를 시각화 할 수 있다는 장점이 있다. 각 축의 거리를 결합하는 방법으로 평균 및 최대, 최소값을 이용할 수 있는데, 향후에는 최대, 최소값을 비교하는 것도 통계적으로 의미있는 일이라 판단된다. 본 연구에서는 함수형 자료의 거리를 바탕으로 다차원 척도법을 접목시켜 시각화한 점에 의의가 있으며, 함수형 자료의 군집화에 확장 및 서명의 위조 탐지 등의 활용에 기대해 본다.

References

Borg, I. and Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Application, 2nd edition*, Springer, New York.

Bubenik, P. and Kim, P. T. (2007). A statistical approach to persistent homology, *Homology, Homotopy and Applications*, **9**, 337–362.

Chung, M. K., Bubenik, P. and Kim, P. T. (2009). Persistence diagrams of cortical surface data, *Information Processing in Medica Imaging (IPMI)*, **21**, 386–397.

Cohen-Steiner, D., Edelsbrunner, H. and Harer, J. (2007). Stability of persistence diagrams, *Discrete and Computational Geometry*, **37**, 103–120.

Delicado, P. (2011a). Dimensionality reduction when data are density functions, *Computational Statistics and Data Analysis*, **55**, 401–420.

Delicado, P. (2011b). *Dimensionality Reduction for Samples of Bivariate Density Level Sets: an Application to Electoral Results* in *Recent Advances in Functional Data Analysis and Related Topics*, ed. Ferraty F., Springer, New York, 71–76.

Edelsbrunner, H. and Harer, J. (2008). Persistent homology - a survey. *Surveys on Discrete and Computational Geometry. Twenty Years Later, eds. J. E. Goodman, J. Pach and R. Pollack, Contemporary Mathematics*, **453**, 257–282.

Edelsbrunner, H., Letscher, D. and Zomorodian, A. (2002). Topological persistence and simplification, *Discrete and Computational Geometry*, **28**, 511–533.

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

Morozov, D. (2008). Homological Illusions of Persistence and Stability. Ph.D. Thesis, Duke University.

Ramsay, R. O. and Silverman, B. W. (2005). *Functional Data Analysis*, Springer, New York.

Zomorodian, A. (2001). *Computing and Comprehending Topology: Persistence and Hierarchical Morse Complexes*, Ph.D. Thesis, University of Illinois, Urbana-Champaign.

Zomorodian, A. and Carlsson, G. (2005). Computing persistent homology, *Discrete and Computational Geometry*, **33**, 249–274.