자연과학 논문집 vol. 26, no. 1, pp. 26-34 (2020) December 2020
Kyung Hee Journal of Natural Sciences

위상적 데이터 분석을 위한 그래픽 사용자 인터페이스 개발

정상만[1], 김경수[2,3,*]

[1]경희대학교 수학과, [2]경희대학교 응용수학과, [3]경희대학교 자연과학종합연구원

Development of graphical user interface for topological data analysis

Sangman Jung[1] and Kyungsoo Kim[2,3,*]

[1]*Department of Mathematics, Kyung Hee University, Seoul 02447*

[2]*Department of Applied Mathematics, Kyung Hee University, Yongin 17104*

[3]*Institute of Natural Sciences, Kyung Hee University, Yongin 17104*

요 약 - 데이터 분석 분야에서 보편적으로 사용되는 방법인 최단 연결 덴드로그램(single linkage dendrogram), 다차원 척도법(multidimensional scaling), k-평균 군집(k-means clustering)과 최근 등장한 위상적 데이터 분석(topological data analysis) 방법을 용이하게 실제 데이터에 적용할 수 있는 MATLAB 기반 그래픽 사용자 인터페이스(GUI) 프로그램을 개발하였다. 개발한 GUI 프로그램을 이용해서 9개 대상의 31개 변수를 100회 측정한 데이터를 분석하여 프로그램의 유용성을 확인하였다.

ABSTRACT - We developed a MATLAB-based graphical user interface (GUI) program for well-known single linkage dendrogram, multidimensional scaling, and k-means clustering and recently developed topological data analysis in data analysis in order to conveniently apply these methods to real data sets. We also provided usefulness of the GUI program by investigating a data set with 31 variables and 100 time steps for 9 subjects.

주 제 어 - GUI 프로그램, 위상적 데이터 분석, 최단 연결 덴드로그램, 다차원 척도법, k-평균 군집
Keywords - GUI program, Topological data analysis, Single linkage dendrogram, Multidimensional scaling, k-means clustering

1. 서론

최근 IT 기술의 빠른 발전으로 인하여 다양한 분야에서 광범위한 데이터를 얻을 수 있게 되면서 데이터 분석에 대한 관심과 중요성이 높아지고 있다. 그러나 데이터의 복잡성이 증가할수록 데이터는 고차원의 비정형 데이터 또는 한정적인 정보만을 갖는 데이터인 경우가 많으므로 기존의 통계적 방법으로는 데이터가 갖는 본질적인 특성을 추출하기 어려운 점이 있다. 위상적 데이터 분석(topological data analysis) 방법은 비교적 최근에 등장한 방법으로 이러한 문제에 대한 한 가지 대안으로 제시되고 있다.[1] 위상적 데이터 분석은 구조의 불변성(invariability)에 대한 연구를 하는 위상수학에 기반한 방법으로, 데이터가 갖는 불변의 위상 구조를 계산적으로 찾아내어 정량화한다.

구체적으로는 데이터가 어떤 모양으로 나타내고, 이 모양과 위상동형(homeomorphic)인 가장 단순한 위상구조를 찾아서 그 위상구조에 대한 정보를 수치적으로 정량화하여 분석함으로써 데이터의 특성을 파악한다.

계산적으로 위상구조를 찾아내고 정량화하는 기법은 지속 호몰로지(persistent homology)를 계산하여 얻어지는 불변량(invariant)인 베티 수(Betti number)를 구하는 것으로 요약된다.[2] 지속 호몰로지는 대수적 위상수학의 심플리셜 호몰로지(simplicial homology) 개념을 도입한 것으로, 데이터가 가질 수 있는 모든 위상구조를 순차적으로 모니터링(monitoring)하면서 각 위상구조의 심플리셜 호몰로지를 지속적으로 계산하는 방법이다. 이때, 모니터링하는 과정을 위상적 데이터 분석에서는 필트레이션(filtration)이라고 부르며, 베티 수는 심플리셜 호몰로지 군의 계수(rank)로 직관적으로는 해당 위상구조가 갖는 차원에서 구멍(hole)의 개수를 의미한다. 결과적으로 위상적 데이터 분석은 주어진 데이터가 가질 수 있는 다양한 위상구조 중 유의미한 위상구조의 특성을 추출하기 위해 데이터가 갖는 모든 위상구조를 순차적으로 조사하여 하나의 위상구조로 수렴할 때까지 반복적으로 각 위상구조에 대한 심플리셜 호몰로지를 계산하고, 이에 대응하는 베티 수를 계산하여 모든 경우에 대한 베티 수 중 유의미한 것들만 뽑아낸다. 이러한 모든 과정은 주로 지속 다이어그램(persistence diagram)과 바코드(barcode)라 부르는 두 시각화 방법으로 요약하여 나타낼 수 있다. 지속 다이어그램은 필트레이션에서 나타나는 위상적 특성의 생성 시점과 소멸 시점을 순서쌍으로 2차원 평면에 나타낸 그림이며, 바코드는 지속 다이어그램의 생성 시점과 소멸 시점을 가로 막대의 시작과 끝으로 표현하여 나타낸 그림이다. 지속 다이어그램은 한 개의 데이터 집합에 대한 위상적 특성을 찾는 문제가 아닌 두 개 이상 데이터 집합의 위상적 특성이 서로 유사하거나 차이가 있는지 확인할 때도 사용되며, 이는 두 데이터 집합에 대한 각각의 지속 다이어그램 사이의 병목 거리(bottleneck distance)를 계산하는 것으로 두 데이터 집합 간 유사성을 정량화하여 나타낼 수 있다.[3]

위상적 데이터 분석은 다양한 분야에서 응용되고 있다.[4-6] Lee et al.은[7] 위상적 데이터 분석 방법에 계층적 군집(hirarchical clustering) 분석 방법 중 최단 연결법(single linkage method)으로 구한 덴드로그램(dendrogram)과[8] Gromov-Hausdorff 거리를[9] 도입하여 인간 뇌의 네트워크 분석에 응용함으로써 서로 다른 세 그룹의 뇌 연결성을 정량화하였다. 최단 연결 덴드로그램은 0차원에서 구멍의 개수, 즉 연결 성분(connected component)을 의미하는 0차 베티 수에 대한 바코드와 최단 연결 덴드로그램의 구성 방식이 동치라는 점을 이용하여 기존의 위상적 데이터 분석 방법으로 얻어지는 위상적 정보와 더불어 위상구조를 이루는 각 변수 간 계층적인 구조에 대한 정보를 얻을 수 있는 장점이 있다. Gromov-Hausdorff 거리는 두 그룹 이상의 덴드로그램의 유사성을 비교하기 위한 것으로 지속 다이어그램의 병목 거리와 마찬가지로 유사성을 측정하는 척도로 사용할 수 있다.

만약 병목 거리와 Gromov-Hausdorff 거리 모두 비교할 그룹의 수가 많은 경우, 각각의 유사도 값을 직관적으로 비교하고 분류하기 어려운 점이 있다. 이를 해결할 방법으로 다차원 척도법(multidimensional scaling)이 이용되고 있다.[10-12] 다차원 척도법은 여러 개체에 대한 상호 유사성을 나타내는 유사성 행렬을 차원 축소하여 저차원의 좌표로 변환하는 기법으로, 보통 2차원 평면 또는 3차원 공간에 각 개체를 점(point)으로 시각화하여 나타내는 기법이다. 다차원 척도법으로 시각화된 결과에서 분류 작업을 할 때는 k-평균 군집법(k-means clustering)을 사용한다. k-평균 군집법은 사용자가 직접 군집 수 k를 적절히 결정해야 한다는 단점이 있으나 scree 플롯을 통한 적당한 k를 결정할 수 있으며, 사용하기 쉽기 때문에 많이 사용되는 군집 방법이다.[13]

한편 위상적 데이터 분석의 다양한 기법을 실질적으로 이용하기 위해 이미 다양한 프로그래밍 언어로 작성된 소프트웨어 패키지가 개발되었다. Python 기반의 GUDHI[14] 과 R 기반의 TDA[15], Java 기반으로 MATLAB과 호환하여 사용할 수 있는 JavaPlex[16]가 대표적이다. 하지만 위상적 데이터 분석 이론 또는 보편적인 데이터 분석의 배경

지식이 많지 않거나 프로그래밍이 익숙하지 않은 경우, 이러한 이론을 직접적으로 응용하여 주어진 데이터를 빠르게 분석하기는 어려울 수 있다.

따라서 본 연구에서는 위상적 데이터 분석의 주요 도구인 바코드, 지속 다이어그램, 병목 거리 계산 및 최단 연결 덴드로그램, Gromov-Hausdorff 거리, 다차원 척도법, k-평균 군집을 용이하게 실제 데이터에 적용할 수 있도록 개발한 그래픽 사용자 인터페이스(graphical user interface, GUI) 프로그램의 데이터 분석 및 응용 방법을 단계적으로 설명하였다. 또한 개발한 GUI 프로그램을 이용해서 9개 대상의 31개 변수를 100회 측정한 데이터를 분석하여 프로그램의 유용성을 제시하였다.

2. GUI 프로그램

이제 여러 가지 데이터 분석 방법을 편리하게 사용하도록 개발한 MATLAB 기반 GUI 프로그램(AppTDA)을 소개한다. 이 프로그램은 위상적 데이터 분석 및 응용에 대한 GUI가 부족한 점을 고려하고, MATLAB 사용자들에게 더 나은 접근성을 제공하기 위해 MATLAB을 기반으로 개발되었다. 하지만 MATLAB에서 기본적으로 제공하는 컴파일러를 이용하면 MATLAB을 설치하지 않는 사용자도 본 프로그램을 사용할 수 있도록 응용프로그램을 만들 수 있으므로, MATLAB 사용자와 비사용자, 나아가 프로그래밍에 익숙하지 않지만 여러 가지 데이터 분석 기법을 응용하려는 사용자까지 고려할 수 있다. 프로그램의 각 분석 기법에 대한 자세한 이론은 본 연구에서 간략하게 언급하였으며, 자세한 내용은 참고문헌에서 확인할 수 있다.

위상적 데이터 분석에서는 분석을 시작하기 전에 주어진 데이터에 위치 정보를 부여하는 형태인 point cloud를 먼저 구한다. 따라서 본 프로그램은 주어진 데이터를 배열의 형태로 불러와서 point cloud를 미리 계산한다. 구체적으로 데이터 집합이 잘 전처리된 3차원 배열 또는 행렬로 주어져있다고 가정한다. 이는 배열의 모든 원소가 수치형(numerical) 데이터 타입을 갖는 데이터 배열 $X \in \mathbb{R}^{N\times M\times P}$임을 의미한다. 데이터 배열 또는 point cloud에서 거리행렬을 만들기 위해 본 연구에서는 거리 공간의 조건을 만족하는 상관계수 기반 거리를 이용하였다.[17] 상관계수 기반 거리는 다음과 같이 정의된다.

X, Y를 각각 중심화 및 정규화된 확률변수라고 하면, 상관계수 기반 거리는 다음과 같이 정의한다.

$$d_C(x,y) = \sqrt{1-|\rho_{xy}|}$$

이때, ρ_{xy}는 $x \in X$와 $y \in Y$의 Pearson 상관계수이다.

데이터 행렬 $X \in \mathbb{R}^{N\times M\times P}$의 각 변수 사이의 상관계수 기반 거리를 계산하면 결과적으로 (i,j) 위치의 원소는 $d_C(x_i,x_j)$이고 대각원소가 모두 0인 대칭행렬 $D_C \in \mathbb{R}^{M\times M\times P}$을 얻는다. 이 거리행렬 D_C를 point cloud로 간주한다.[18]

GUI 프로그램 AppTDA의 데이터 분석 흐름은 간략하게 다음과 같은 과정으로 요약된다.

(1) 데이터의 각 변수 사이에 적당한 거리를 정의하여 거리행렬을 구한다.
(2-1) 거리행렬에 대한 지속 호몰로지를 계산하여 지속 다이어그램을 구한다.
(2-2) 거리행렬에 대한 최단 연결 덴드로그램을 계산하여 행렬의 형태로 저장한다.
(3-1) 데이터가 두 개 이상의 그룹으로 나누어지는 경우 지속 다이어그램 간의 병목 거리를 계산한다.
(3-2) 최단 연결 행렬 사이의 Gromov-Hausdorff 거리를 계산한다.
(4) 병목 거리와 Gromov-Hausdorff 거리 모두 다차원 척도법을 통해 시각화한다.
(5) k-평균 군집을 통해 시각화된 개체들의 군집을 찾는다.

실제로 프로그램에서 데이터를 불러올 경우, 분석 도구들을 사용하기 전에 (1)의 거리행렬과 (3-2)의 최단 연결 행렬 사이의 Gromov-Hausdorff 거리를 미리 계산하여 저장한다. 이 과정은 그림 1의 flowchart와 같이 나타낼 수 있다.

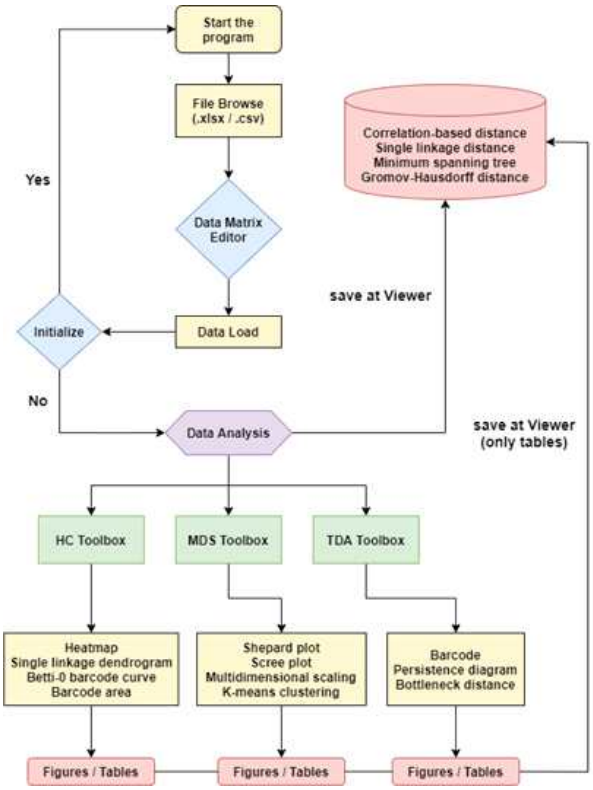

그림 1. AppTDA flowchart

AppTDA는 크게 두 개의 패널로 구성되어 있다 (그림 2). 첫 번째는 데이터를 편집하고 불러오기 위한 'Data Loading' 패널이고 두 번째는 실질적인 데이터 분석을 위한 'Data Analysis' 패널이다.

Data Loading 패널에서는 Browse 버튼을 눌러 팝업창 안에서 데이터가 위치한 화일 경로를 탐색하여 선택할 수 있으며, Browse 작업이 끝나고 나면 데이터 배열이 갖는 데이터 크기를 자동으로 계산하여 불러온다. 사용자는 데이터에서 일정 범위만 불러오고 싶을 때 Data Matrix Editor에서 Yes를 선택하고 각 rows, columns, pages를 조절하여 불러올 수 있다. 데이터 크기를 편집하는 칸 왼쪽 아래에 Scaling and Centering 체크박스를 선택하면 데이터 배열에 중심화와 정규화를 적용할 수 있다. 최종적으로 Load 버튼을 누르면 편집이 완료된 데이터 배열의 상관계수 기반 거리행렬, 최단 연결 행렬, 최단 연결 행렬에 따른 최소생성 트리, Gromov-Hausdorff를 계산하여 Viewer의 우측 하단에 있는 창에 표시한다. 계산된 값은 모두 Data Loading 패널 아래의 Viewer에서 확인할 수 있으며, xlsx 파일 형식으로 저장할 수 있다.

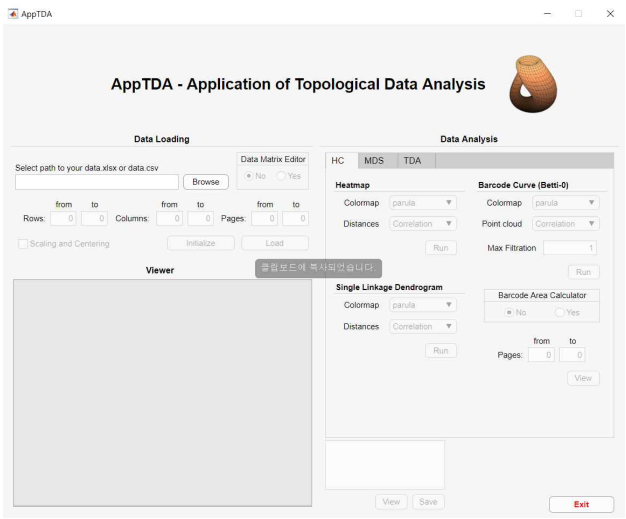

그림 2. AppTDA 실행 화면

만약 새로운 데이터를 불러와서 분석을 하거나 기존에 편집한 결과를 바꾸는 경우 Initialize 버튼을 누르면 결과를 초기화할 수 있도록 설계하였다. 프로그램의 모든 기능은 반드시 선행 요구사항이 완료되기 전에는 사용할 수 없도록 블라인드 처리가 되어 있다.

Data Analysis 패널은 세 개의 탭(Tab)으로 구성되어 있으며, 패널의 좌측부터 HC, MDS, TDA이다. HC 탭은 계층적 군집 분석 도구함, MDS 탭은 다차원 척도법 도구함, TDA 탭은 위상적 데이터 분석 도구함을 의미한다. HC 탭은 세 가지 기능을 포함하고 있는데, 첫 번째로 Heatmap 시각화 도구, 두 번째로 최단 연결 덴드로그램, 마지막으로 0차 베티 수에 대한 바코드를 곡선으로 표현하는 기능이다 (그림 3). 각 기능의 Colormap은 시각화 결과에서 색의 표현을, Distances는 거리행렬을 선택할 수 있도록 하였다. 바코드 곡선을 구하는 기능에서 Point cloud는 Distances와 동일하며, 바코드의 최대 필트레이션 값을 조절할 수

있도록 Max Filtration 칸이 존재한다. 또한 바코드 곡선 아래의 넓이를 구할 수 있는 기능이 있으며 이는 Barcode Area Calculator의 Yes를 선택하여 사용할 수 있다. 최종적으로 각 기능에서 원하는 설정을 선택한 후 Run을 누르면 기능이 실행된다.

MDS 탭은 다차원 척도법을 적용하기 위한 도구함으로, Shapard 플롯, scree 플롯, 다차원 척도법, k-평균 군집을 포함한다 (그림 4). Shepard 플롯은 다차원 척도법으로 차원 축소된 결과값의 적합도(goodness of fit)를 판정하고자 할 때 쓰이는 기법으로, 단순히 입력값과 결과값을 2차원 평면에 산포도로 나타내어 입력값과 결과값의 오차를 비교한다. AppTDA는 Shapard 플롯과 더불어 산포도의 Pearson 상관계수도 함께 표시해준다. scree 플롯은 k-평균 군집의 k를 결정하기 위한 것으로, k-평균 군집 시 해당 군집 수의 군집 내 제곱합(the within-cluster sum of squares, WSS)의 크기가 작을수록 적합함을 이용하여 가로축에는 k의 수, 세로 축에는 WSS 값을 적용하여 군집 수 k를 결정한다. Shepard 플롯, scree플롯, 다차원 척도법 기능의 선택 항목 중 Dissimilarity는 앞서 언급한 Distances를 의미하고 Dimension은 다차원 척도법으로 차원 축소할 차원의 수를 의미한다. Criterion은 다차원 척도법에서 사용할 손실함수(loss function)을 의미한다. 다차원 척도법을 실행하기 전 k-평균 군집을 적용하려는 경우에는 K-means Clustering 박스에서 Yes를 선택하면 Number of K가 활성화되면서 군집 수를 결정할 수 있다. 계산된 다차원 척도법 결과와 적합도는 자동으로 Viewer 우측 하단의 창에 저장되어 결과 확인 및 저장이 가능하다.

TDA 탭은 위상적 데이터 분석의 주요 도구인 바코드, 지속 다이어그램, 다중 데이터에 대한 유사성 비교를 위한 병목 거리 계산 기능을 포함하고 있다 (그림 5). 각 기능은 JavaPlex 패키지를 참조하여 구현하였다. 본 프로그램의 병목 거리는 상관계수 기반 거리만 point cloud로 사용할 수 있으므로 point cloud를 선택하지 않는다. 선택 항목 중 Betti Number는 베티 수의 차원을 의미하며 0 차원부터 최대 2차원까지의 결과를 확인할 수 있도록 하였다. 바코드와 지속 다이어그램은 Run 버튼을 누르면 주어진 데이터 배열의 page 수만큼 바코드와 지속 다이어그램을 계산하여 출력한다. 병목 거리의 경우는 page×page 크기의 병목 거리 행렬을 Viewer와 모든 탭의 Distances, Point cloud, Dissimilarity 항목에 추가된다. 그리고 프로그램 우측 최하단의 Exit 버튼을 누르면 프로그램을 종료할 수 있도록 하였다.

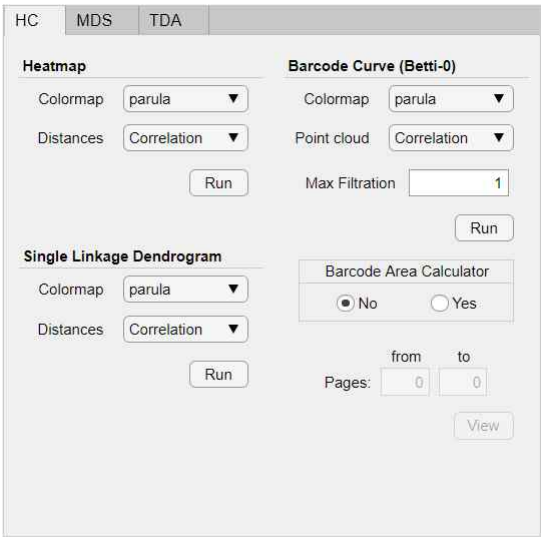

그림 3. HC 탭과 내부 기능

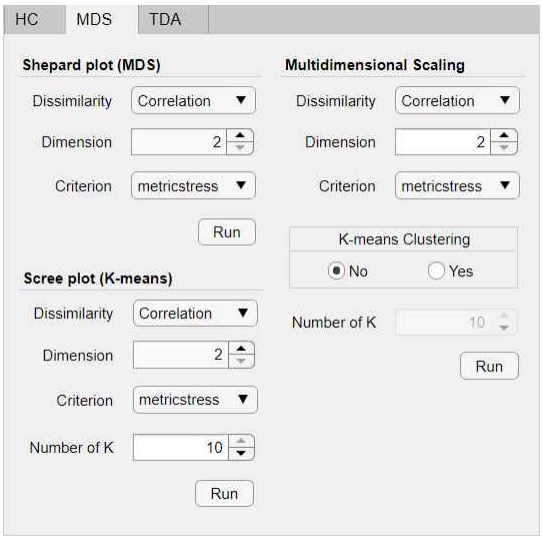

그림 4. MDS 탭과 내부 기능

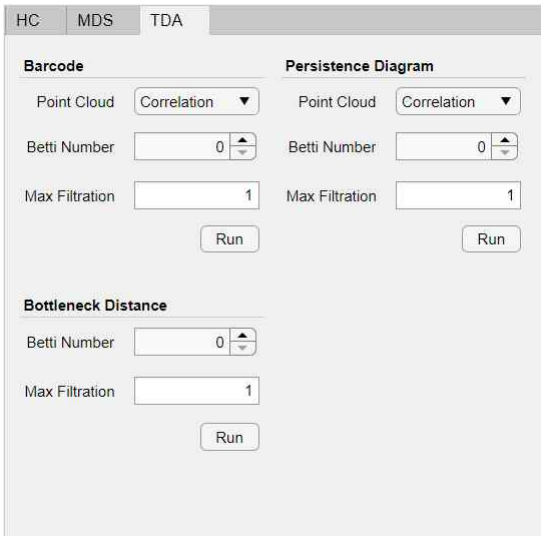

그림 5. TDA 탭과 내부 기능

3. 데이터 분석 결과

9개 대상의 31개 변수를 100회 측정하여 엑셀 화일로 저장한 데이터, 즉 $100 \times 31 \times 9$ 크기의 3차원 배열을 테스트 데이터로 개발한 GUI 프로그램 AppTDA를 이용하여 분석하고 프로그램의 유용성을 확인하였다.

AppTDA를 실행하고 테스트 데이터를 불러오면 (그림 6), 자동적으로 9개 대상의 31개의 변수 사이의 상관계수 기반 거리 행렬 D_C을 계산한다. 먼저 각 변수 사이의 비유사도를 heatmap으로 시각화하여 우선 고려할 변수가 무엇인지 확인한다 (그림 6). 31개 변수 중 비유사도가 큰 부분은 파란색 계열로, 비유사도가 작은 부분은 노란색 계열로 나타난다.

다음으로 위상적 특성을 분석하기 위해 D_C의 바코드와 지속 다이어그램을 확인한다. 그림 7은 9개 대상 중 2번 대상에 대한 0차, 1차, 2차 베티 수에 대한 바코드와 지속 다이어그램을 나타낸다. 지속 다이어그램에서 1차 베티 수에 대하여 유의미한 위상적 특성이 두 개가 존재함을 확인할 수 있다.

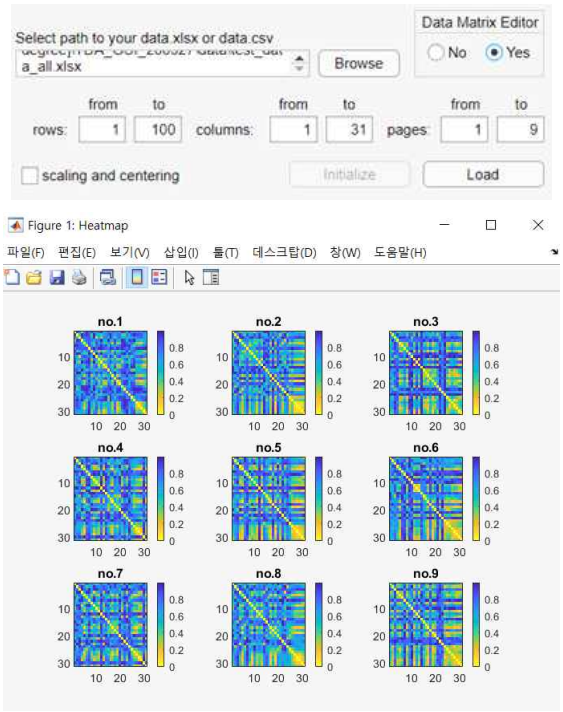

그림 6. 데이터 불러오기 화면과 관계수-기반 거리에 대한 heatmap

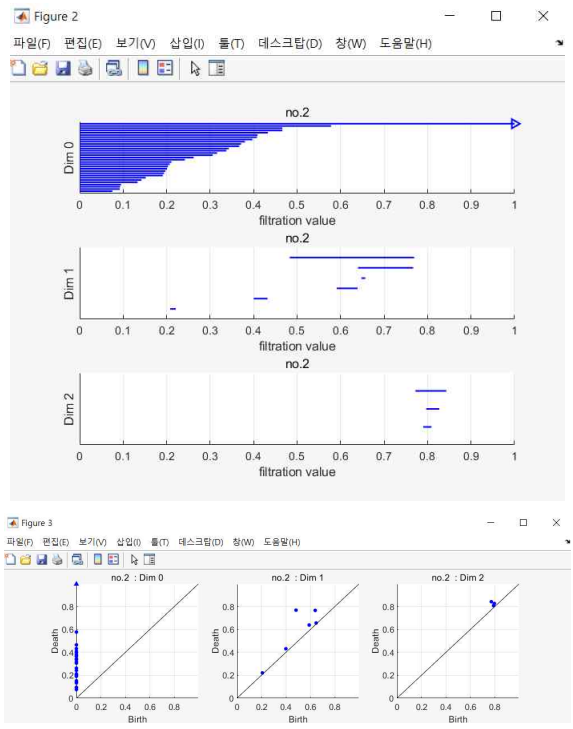

그림 7. 2번 대상의 바코드와 지속 다이어그램

또한 0차 베티 수에 대한 바코드를 곡선 형태로 변환하여 확인할 수도 있다 (그림 8). 9개 대상에 대한 바코드 곡선을 비교하여 개체 사이의 차이를 확인할 수 있으며, 각 바코드 곡선 아래 영역의 넓이를 계산함으로써 그 차이를 정량화할 수 있다.

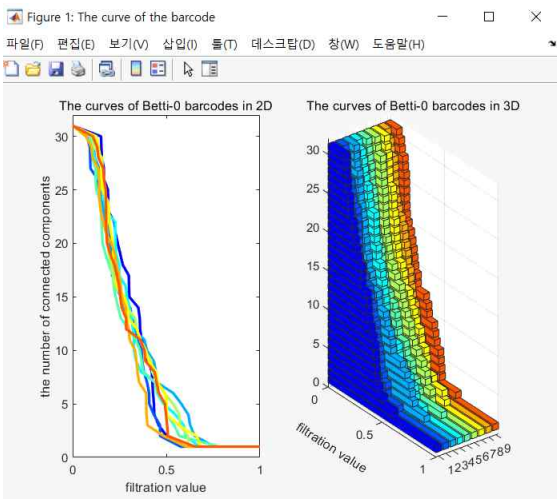

그림 8. 0차 베티 수에 대한 바코드 곡선

0차 베티 수에 대한 바코드에서 파악하기 어려운 각 변수 사이의 계층적 관계는 최단 연결 덴드로그램을 통해 분석할 수 있다. 그림 9는 덴드로그램 세로 축의 1부터 31까지에 해당하는 변수들은 가로 축의 필트레이션 값에 따라 서로 어떠한 관계로 연결되어 있는지를 보여준다.

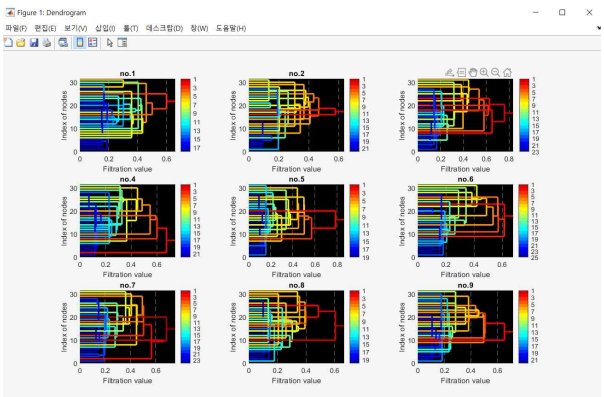

그림 9. 9개 대상에 대한 최단 연결 덴드로그램

9개 대상 각각의 위상적, 계층적 특성을 모두 파악하고, 대상 사이의 유사도를 확인하기 위해 지속 다이어그램을 비교하는 병목 거리행렬과 최단 연결 덴드로그램을 비교하는 Gromov-Hausdorff 거리 행렬을 구한다. 그리고 두 거리행렬을 비유사도 행렬로 가정하고 다차원 척도법을 적용한다. 다차원 척도법의 손실함수를 결정하는 것은 Shepard 플롯을 이용할 수 있는데, 차원을 높이거나 손실함수를 바꾸는 경우 Shepard 플롯의 입력과 출력으로 구성된 순서쌍이 직선 $y = x$에 근접하는지, 고르게 분포하는지 등을 확인해야 한다. 그림 10은 3차원에서의 stress 손실함수를 적용하였을 때 적합도 결과를 보여준다. 여기서 D, C, P는 각각 차원, 손실함수, Pearson 상관계수를 의미한다.

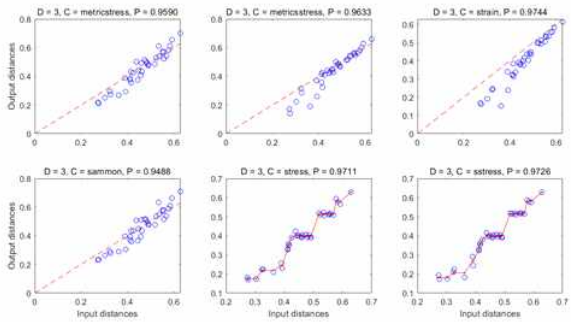

그림 10. Gromov-Hausdorff 거리의 3차원 다차원 척도법 결과에 대한 Shepard 플롯

Shepard 플롯을 통해 손실함수를 stress로, 3차원으로 차원 축소를 결정하고, k-평균 군집 알고리즘을 이용하기 위해 이제 군집 수 k를 정한다. scree 플롯을 통해 elbow가 생기는 지점인 k = 3을 군집 수로 정한다 (그림 11). 최종적으로 다차원 척도법을 적용하여 그림 12를 얻을 수 있다. 그림 12에서 대상 4, 5, 7은 Group A로, 대상 1, 8, 9는 Group B로, 대상 2, 3, 6은 Group C로 군집되는 것을 시각적으로 확인할 수 있다. 서로 다른 그룹에 속한 대상들은 Gromov-Hausdorff 거리 만큼의 비유사도를 갖는다고 할 수 있겠다. 병목 거리를 이용하는 경우도 같은 방법으로 분석할 수 있다.

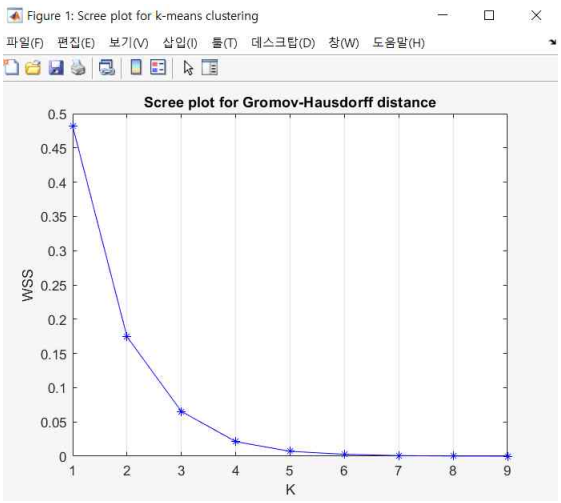

그림 11. Gromov-Hausdorff 거리의 3차원 다차원 척도법 결과에 대한 scree 플롯

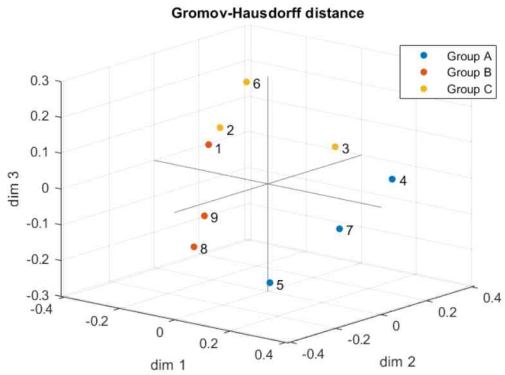

그림 12. Gromov-Hausdorff 거리의 다차원 척도법 및 k-평균 군집 결과

4. 결론

본 연구에서는 데이터 분석 분야에서 보편적으로 사용되는 방법인 최단 연결 덴드로그램, 다차원 척도법, k-평균 군집과 최근 등장한 위상적 데이터 분석 방법을 실제 데이터에 적용할 수 있는 MATLAB 기반 GUI 프로그램의 데이터 분석 및 응용 방법을 단계적으로 설명하였다. 이 프로그램은 직접적인 알고리즘 구현이나 프로그래밍 작업을 거치지 않고 쉽게 사용할 수 있으며, 9개 대상의 31개 변수를 100회 측정한 데이터를 분석하여 프로그램의 유용성을 확인하였다. 향후 업데이트를 통하여 위상적 데이터 분석의 다양한 최신 분석 도구를 추가하고, 효율성을 향상시키면, 더욱 편리한 데이터 분석 프로그램이 될 것으로 기대된다.

참고문헌

1. Carlsson, G. (2009). Topology and data. Bulletin of the American Mathematical Society, 46(2), 255-308.
2. Edelsbrunner, H., Harer, J. (2008). Persistent homology - a survey. Contemporary Mathematics, 453, 257-282.
3. Cohen-Steiner, D., Edelsbrunner, H., Harer, J. (2007). Stability of persistence diagrams. Discrete & Computational Geometry, 37(1), 103-120.
4. Ferri, M. (2017). Persistent topology for natural data analysis - a survey. In Towards Integrative Machine Learning and Knowledge Extraction, Springer, 117-133.
5. Pun, C. S., Xia, K., Lee, S. X. (2018). Persistent-Homology-based machine learning and its applications - a survey. arXiv:1811.00252.
6. Lee, H., Chung, M. K., Kang, H., Kim, B. N., Lee, D. S. (2011). Discriminative persistent homology of brain networks. In IEEE International Symposium on Biomedical Imaging, 841-844.
7. 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.
8. Carlsson, G., Mémoli, F. (2010). Characterization, stability and convergence of hierarchical clustering methods. The Journal of Machine Learning Research, 11, 1425-1470.

9. Mémoli, F. (2008). Gromov-Hausdorff distances in Euclidean spaces. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 1-8.
10. 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. F1000 Research, 7.
11. Costa, J. P., Škraba, P. (2015). A topological data analysis approach to the epidemiology of influenza. In SIKDD15 Conference Proceedings.
12. Hajij, M., Jonoska, N., Kukushkin, D., Saito, M. (2018). Graph-based analysis for gene segment organization in a scrambled genome. arXiv:1801.05922.
13. Han, J., Pei, J., Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
14. Maria, C., Boissonnat, J. D., Glisse, M., Yvinec, M. (2014). The Gudhi library: Simplicial complexes and persistent homology. In International Congress on Mathematical Software, Springer, 167-174.
15. Fasy, B. T., Kim, J., Lecci, F., Maria, C. (2014). Introduction to the R package TDA. arXiv:1411.1830.
16. Adams, H., Tausz, A. (2015). Javaplex tutorial. Retrieved from http://goo.gl/5uaRoQ.
17. 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.
18. Ghrist, R. (2008). Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society, 45(1), 61-75.