The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.6.999
26(6), 999–1007

Moving Data Pictures

Myung-Hoe Huh[a,1]

[a]Department of Statistics, Korea University

(Received October 8, 2013; Revised November 16, 2013; Accepted November 21, 2013)

Abstract

This research shows several types of moving pictures from the data: 1) the word cloud of Korean texts, 2) the heat map of $n \times p$ matrices, 3) the moving image of $p \times p$ scatterplot matrix, 4) the local projective display of k clusters (Huh and Lee, 2012). Moving pictures may reveal the hidden information and beauty of the datasets and ignite the curiosity of information consumers. Video files are attached.

Keywords: Data visualization, dynamic graphics, moving pictures, word cloud, heat map, scatterplot matrix, k-means clustering.

1. 연구 내용과 방법

3차원 산점도 회전(3D scatterplot rotation)과 이를 확장한 G고비(GGobi)는 움직이는 데이터 그림의 대표적 보기이다 (http://www.ggobi.org; Cook과 Swayne, 2007). 움직이는 그림(moving pictures), 즉 동영상(video image)은 현대사회에서 가장 강력한 미디어 도구이고 정보문화이므로, 데이터 시각화(data visualization) 분야에서도 움직이는 형태의 그림에 관심을 둘 필요가 있다.

이 연구는 1) 한국어 텍스트의 단어 구름, 2) $n \times p$ 행렬의 시각화, 3) $p \times p$ 산점도 행렬의 동영상 버전, 4) k개 개체 군집의 동적 시각화 등에 적용될 수 있는 움직이는 데이터 그림을 제안 한다.

이 연구에서 사용된 계산 도구는 오픈소스의 통계 언어소프트웨어인 R이다. R은 매우 빠른 속도로 고질의 그래프를 만들어 내므로 수십 수백 개의 그래프를 연이어 전시함으로써 동영상 효과를 연출할 수 있다. 오히려 그래프 화면들이 너무 빨리 넘어가지 않도록 `Sys.sleep()` 함수를 써서 실행 시간을 지연시킬 필요가 있다. Figure 1.1은 2개 자료 점의 움직임을 보여주는 1개의 보기이다. 그 동영상은 1초당 4개의 그래프 화면을 보여준다 (참고: `avi` 파일은 마이크로소프트사가 만든 멀티미디어 형식을 따르는 비디오 파일로서 MS 미디어플레이어 등에서 재생이 가능하다.).

움직이는 데이터 그림은 데이터에 내재된 시각적 아름다움을 보여주어 그 자체로 미술적 가치가 있다. 또한 정보 소비자들이 데이터에 흥미를 갖게 하고 데이터에 대한 포괄적 이해를 도울 수 있다.

2. 한국어 텍스트의 단어 구름

단어 구름(word cloud)은 m개의 단어를 출현 빈도에 비례하는 크기로 표출한 그래프이다. 인터넷 사이트 wordle.net은 사용자의 텍스트를 즉석에서 단어 구름으로 만들어준다. 다만, 2013년 현재 한국어 텍스트는 처리되지 않는다.

[1]Professor, Department of Statistics, Korea University, Anam-Dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea. E-mail: stat420@korea.ac.kr

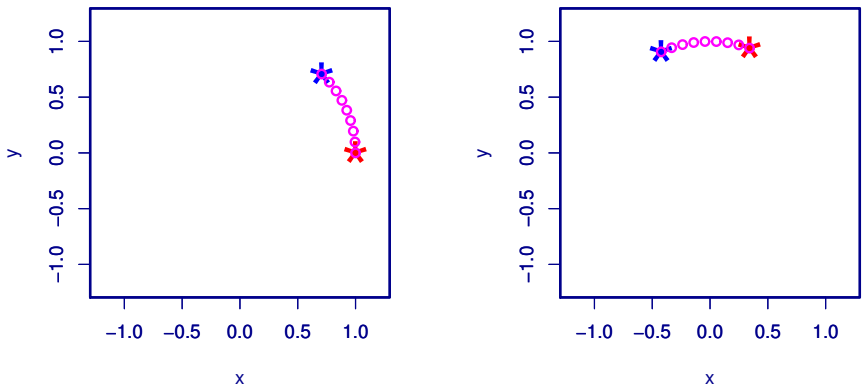

Figure 1.1. Moving pair of stars (Video File: 'moving picture demo.avi')

Figure 2.1. Word cloud from e-mail essay (Video File: 'wordcloud email-essay.avi')

R의 `KoNLP` 팩키지는 한국어 텍스트에 대한 형태소 분석(morphological analysis) 기능을 가지고 있어 텍스트에 포함된 명사들을 추출해낸다. 명사는 텍스트에 포함된 내용을 담으므로, 명사 리스트와 출현 빈도로부터 텍스트 내용을 대략적으로 알 수 있다. 단어 구름은 이를 시각화함으로써 잠재적 독자들에게 경제적이고 효과적인 요약을 제공한다.

Figure 2.1은 허명회의 2011년 고려대학교 정경대학 신문 에세이 "e-메일의 기본 격(格)"(200자 원고지 9매 분량)의 단어 구름이다. 각 단어는 출현 빈도에 비례하는 크기로 찍히므로 '메일', '학생', '교수' 등이 주요 단어임을 한 눈에 알아볼 수 있다.

완성 화면이 보여지기 전에, 단어 구름이 생성되는 모습을 동영상으로 제시하면 어떨까? 텍스트 읽기가 순차적 과정이라는 점을 존중하자는 것이다. 텍스트에서 읽힌 단어가 명사로 판명되는 대로 화면에 표출시키고 일정 시간 머물게 되므로 각 단어는 텍스트 내 빈도의 횟수로 동영상 화면에 나타난다. 전체 텍스트가 스캔되고 나면 완성된 그래프가 제시된다.

Figure 2.2. Word cloud from the story of rabbit and tortoise (Video File: 'wordcloud rabbit-story.avi')

Figure 2.2는 "토끼와 거북이의 두 번째 경기"에 대한 움직이는 단어 구름이다. 동영상을 보는 과정에서 텍스트의 내용이 무엇인지 추측하게 되고 마지막으로 제시된 요약에서 추측의 옳고 그름을 알 수 있게 된다. 이것은 그냥 결과만 보는 것과는 흥미의 격이 다르다.

3. 열(熱) 지도

열 지도(heat map)는 수치 행렬 X의 칸 값들을 일련의 컬러로 시각화한 그래프이다. 관건은 전체 그래프가 의미 있는 패턴을 보이도록 n개 행과 p개 열을 재배열하는 데 있다. 재배열 알고리즘에는 여러 가지가 있으며 재배열 결과는 다를 수 있다. 행렬 시각화(matrix visualization)에 대한 학술적 검토와 알고리즘은 Wu 등 (2008)에서 찾을 수 있다. 이 절에서는 다음 2종의 알고리즘을 쓸 것이다.

알고리즘 1. 행과 열의 교대 재배열(alternating rearrangement)

0) 행 간 거리가 가장 큰 2개 행을 찾아 하나를 행 1에 넣고 그 행으로부터의 거리 순서로 나머지 행들은 재배열 한다. 행 $i\,(=1,\ldots,n)$에 점수 $i-(1+n)/2$를 부여하고 열 $j\,(=1,\ldots,p)$에 점수 $j-(1+p)/2$를 부여한다.

1) 각 열에 대하여 열 벡터와 행 점수 간 내적을 산출하고 그 순서에 따라 열을 재배열 한다.

2) 각 행에 대하여 행 벡터와 열 점수 간 내적을 산출하고 그 순서에 따라 행을 재배열 한다.

3) 단계 1과 단계 2를 더 이상의 변화가 없을 때까지 반복한다.

알고리즘 2. Hurley (2004)의 끝 잇기(endlink)

1) n개 행(개체) 간 거리에 끝 잇기 알고리즘을 적용한다. Hurley의 알고리즘은 개체 간 거리의 총합이 최소가 되도록 행들을 한 줄로 잇는다.

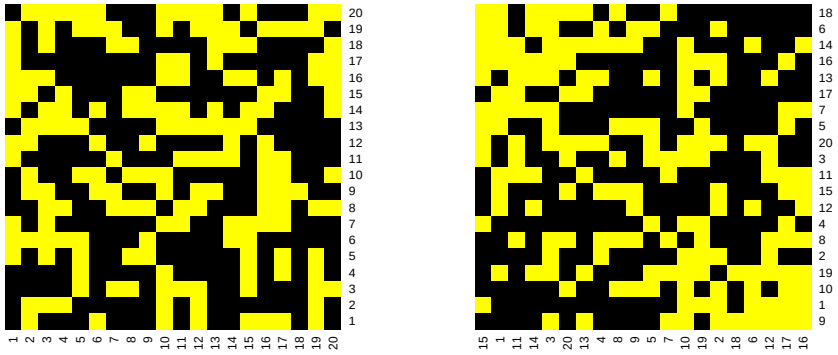

Figure 3.1. Changing images of the 20×20 random matrix, at the start (Left) and the end (Right)

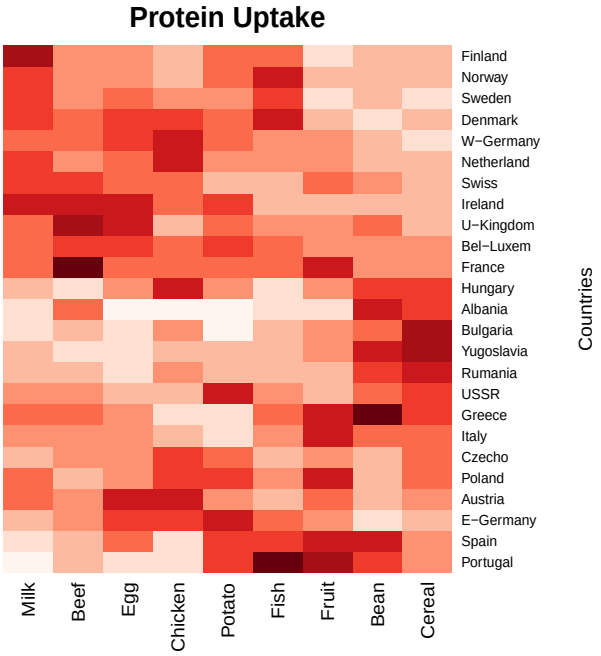

Figure 3.2. Heat map of the European 25 countries protein uptake data

2) p개 열(변수) 간 상관에 끝 잇기 알고리즘을 적용한다. Hurley의 알고리즘은 변수 간 상관의 총 합이 최대가 되도록 열들을 한 줄로 잇는다.

Figure 3.1의 왼쪽 것은 각 칸이 독립적으로 0.5의 확률로 0 또는 1의 값을 취하는 20×20 임의 행렬에 대한 열(熱) 지도이고 오른쪽 것은 그 행렬에 알고리즘 1을 적용하여 행과 열이 재배열된 최종 행렬에 대한 열 지도이다. 그 과정에서 단계 1과 단계 2가 13회 반복되었는데, 재배열의 과정을 동영상으로 보임으로써 생동감을 유도한다 (동영상 파일: heatmap matrix 20×20 alternating.avi). 최종 열 지도에서 1의 값을 갖는 검정색 칸들이 좌하-우상의 대각선상에 집중되어 있음을 볼 수 있다.

Figure 3.2는 유럽 25개 나라 국민의 단백질 섭취원 자료에 열(列) 표준화와 알고리즘 2에 의한 행과 열의 재배열이 적용된 열 지도이다. 적용된 컬러링 팩키지는 `RColorBrewer`이다 (http://colorbrewer2.org). 행과 열의 재배열 과정에서 동영상 열 지도가 생성된다 (동영상 파일: heatmap protein endlink.avi). 최종 열 지도에서 S자형의 패턴을 볼 수 있다.

4. 산점도 행렬

1개의 종속변수와 이에 대한 p개의 설명변수로 구성된 다변량 자료를 탐색하는 상황에서 통상적으로 활용되는 도구는 산점도 행렬(scatterplot matrix)이다. 그런데 p가 3∼4를 넘어가는 경우 플롯 1개의 크기가 너무 작게 되는 문제가 생긴다. 대안으로서 다음과 같은 움직이는 산점도를 생각하기로 한다.

종속변수 Y와 설명변수 $X_1, \ldots, X_p$ 간 관계의 탐색:

0) $X_1, \ldots, X_p$에 끝 잇기 알고리즘을 적용하여 순서화한다. 그것을 $X_{j_1}, \ldots, X_{j_p}$로 칭한다.

1) Y와 X_{j_1,θ,j_2} 간 산점도를 연이어 만든다. 여기서 X_{j_1,θ,j_2}는 $X_{j_1}\cos(\theta)+X_{j_2}\sin(\theta)$이고 θ는 0에서 $\pi/2$까지 변한다. 이에 따라 Y와 X_{j_1} 간 산점도에서 출발하여 Y와 X_{j_2} 간 산점도에 귀착되는 연속적 이미지의 동영상이 생성된다.

2) Y와 X_{j_2,θ,j_3} 간 산점도를 연이어 만든다. 여기서 X_{j_2,θ,j_3}는 $X_{j_2}\cos(\theta)+X_{j_3}\sin(\theta)$이고 θ는 0에서 $\pi/2$까지 변한다. 이에 따라 Y와 X_{j_2} 간 산점도에서 출발하여 Y와 X_{j_3} 간 산점도에 귀착되는 연속적 이미지의 동영상이 생성된다.

⋮

p) Y와 X_{j_{p-1},θ,j_p} 간 산점도를 연이어 만든다. 여기서 X_{j_{p-1},θ,j_p}는 $X_{j_{p-1}}\cos(\theta)+X_{j_p}\sin(\theta)$이고 θ는 0에서 $\pi/2$까지 변한다. 이에 따라 Y와 $X_{j_{p-1}}$ 간 산점도에서 출발하여 Y와 X_{j_p} 간 산점도에 귀착되는 연속적 이미지의 동영상이 생성된다.

설명변수 $X_1, \ldots, X_p$ 간 관계의 탐색:

0) 각 $j\,(=1,\ldots,p)$에 대하여 다음 단계의 작업을 한다.

1) 변수 X_j를 종속변수로 간주하고 나머지 변수들을 설명변수로 간주하여 앞의 알고리즘을 적용한다. 즉, j번째 변수를 제외한 나머지 $p-1$개 변수들을 끝 잇기 알고리즘으로 배열하여 그 순서대로 X_j와의 산점도를 그린다.

Figure 4.1은 330일에 걸쳐 측정된 오존(ozone, 종속변수)과 이와 관련된 8개(= p) 공변량 간 산점도 행렬이다 (Breiman과 Friedman, 1985). 끝 잇기 알고리즘을 적용한 결과 $X_5 - X_7 - X_1 - X_6 - X_3 - X_8 - X_2 - X_4$로 정렬되었다.

Figure 4.2는 Y와 $X_{5,\theta,7}$ 간 산점도 2개이다 ($\theta = 0.2\cdot\pi/2, 0.8\cdot\pi/2$; X_5는 Hgt, X_7은 InvTmp). 종속변수 Y와 설명변수 $X_5, X_7, \ldots, X_2, X_4$ 간 산점도의 동영상은 종속변수와 설명변수가 정(正) 관계에서 출발하여 부(否) 관계로 종착되는 데이터 구름의 변화를 보여준다.

Figure 4.3은 (X_4, X_1) 간 산점도에서 (X_2, X_1) 간 산점도로 이동하는 도중의 두 플롯으로 (X_1: Temp, X_2: InvHt, X_4: Vis) 왼쪽 것은 출발점 가까이에서 찍힌 산점도이고 오른쪽 것은 도착점 가까이에서 찍힌 것이다. 오른쪽 플롯에서 어렴풋하게 2개의 군집이 형성된 것을 볼 수 있는데, 이와 같이 $p(\geq 3)$개 설명변수 간 산점도의 동영상으로 흥미로운 패턴의 포착 기회를 잡을 수 있다.

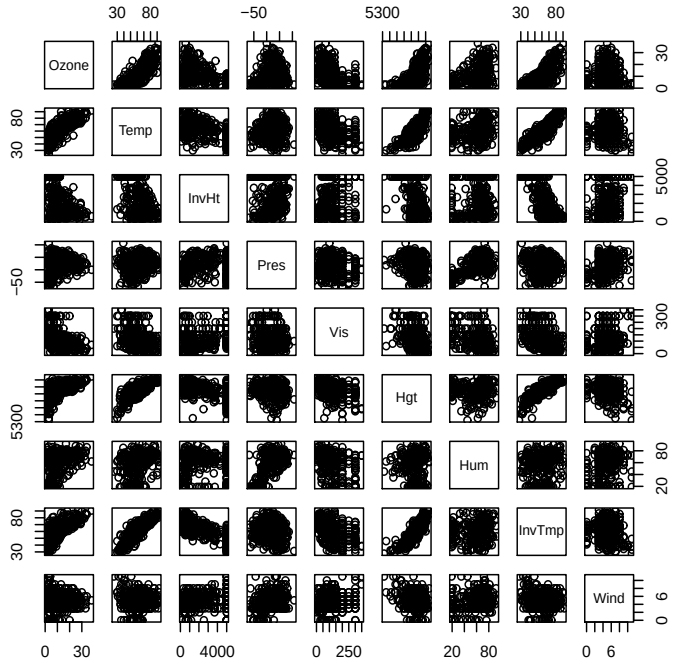

Figure 4.1. Scatterplot matrix of the Ozone Data (Breiman and Friedman, 1985)

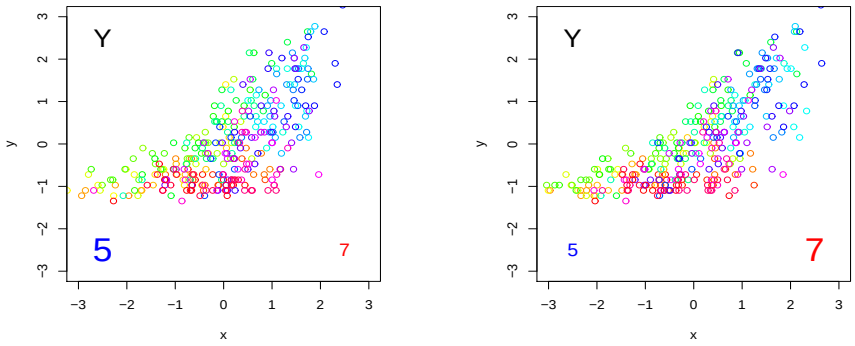

Figure 4.2. The Relationship between Y Variable and X Variables (Video File: 'scatterplot ozone yx.avi')

5. k개 군집의 시각화 표현

k-평균 군집화는 n개의 p변량 개체를 k개 군집으로 묶어낸다. 이 외 다른 군집화 방법들도 데이터세트 내 개체들을 다수의 군집으로 묶어낸다. 그러나 이제까지는 k개 군집을 효과적으로 시각화하기 어려웠다. 최근 Huh와 Lee (2012)는 각 군집에 포커스를 둔 국소적 사영 전시 기법을 제안하였는데, 이 방법에서는 군집별로 저차원 사영에 쓰는 기저 벡터가 다르다.

군집 j의 2차원 국소적 사영이 p차원 단위벡터 $\mathbf{u}_j$와 $\mathbf{v}_j$에 의하여 주도된다고 하자 ($j = 1, \ldots, k$). 이에 따라 군집 j의 2차원 국소적 사영 플롯은 $(X\mathbf{u}_j, X\mathbf{v}_j)$의 n개 행으로 구성되고 플롯에서 군집 j의 개체

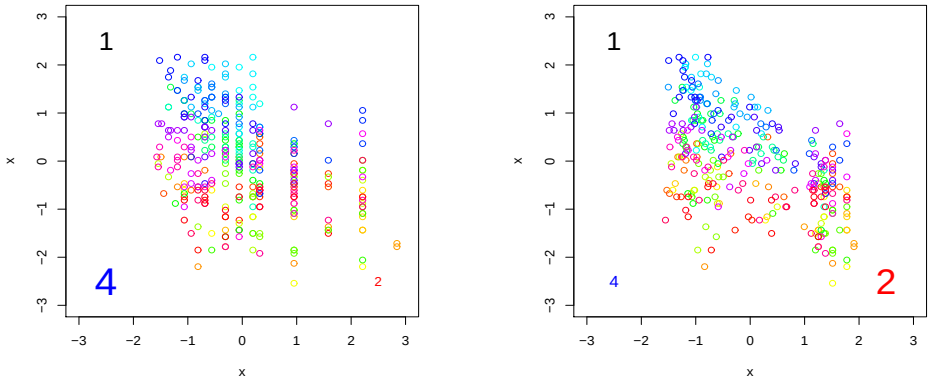

Figure 4.3. The Relationship between X Variables (Video File: 'scatterplot ozone xx.avi')

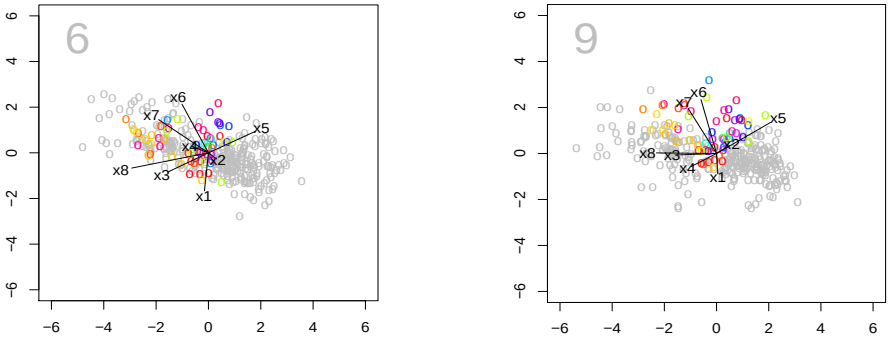

Figure 5.1. Transition from Cluster 1 to Cluster 2 Projections of the Italian Olive Oil Data

들은 전면 기호로, 그 외 군집의 개체들은 후면 기호로 표출된다. 군집 j 플롯의 다음이 군집 j' 플롯이고 이 플롯이 $\mathbf{u}_{j'}$과 $\mathbf{v}_{j'}$에 의하여 주도되는 경우, 다음 두 단위벡터를 사영벡터로 함으로써 군집 j 플롯과 군집 j' 플롯을 보간(補間)할 수 있다:

$$\mathbf{u}^*_{j,\theta,j'} = \frac{\mathbf{u}_{j,\theta,j'}}{\|\mathbf{u}_{j,\theta,j'}\|}, \qquad \mathbf{v}^*_{j,\theta,j'} = \mathbf{v}_{j,\theta,j'} - \frac{< \mathbf{u}_{j,\theta,j'}, \mathbf{v}_{j,\theta,j'} >}{\|\mathbf{u}_{j,\theta,j'}\|^2}\mathbf{u}_{j,\theta,j'},$$

여기서 $\mathbf{u}_{j,\theta,j'} = \mathbf{u}_j \cos(\theta) + \mathbf{u}_{j'} \sin(\theta)$이고 $\mathbf{v}_{j,\theta,j'} = \mathbf{v}_j \cos(\theta) + \mathbf{v}_{j'} \sin(\theta)$이다. 이에 따라 군집 j와 군집 j' 간 보간 플롯은 $(X\mathbf{u}^*_{j,\theta,j'}, X\mathbf{v}^*_{j,\theta,j'}/\|\mathbf{v}^*_{j,\theta,j'}\|)$의 n개 행으로 구성된다 $(0 \le \theta \le \pi/2)$.

Huh와 Lee (2012)는 국소적 사영 방법의 수치 예로서 이탈리아 올리브油 자료에 $k = 4$인 k-평균 군집화를 적용하고 끝 잇기 알고리즘으로 군집 플롯들을 순차적으로 제시한 바 있다. 이 자료는 323종(= n)의 올리브油에 대한 8개(= p) 지방산 성분의 측정값으로 구성되어 있다.

Figure 5.1은 $k = 4$인 k-평균 군집화에서 군집 1 사영에서 군집 2 사영으로 넘어가는 10개 중간 과정에서 6번째 화면과 9번째 화면이다. 화면에서 군집 2 개체들이 컬러로 마킹되어 있다. 이 사례의 군집 시각화 동영상은 군집 1 → 군집 2 → 군집 4 → 군집 3의 전체 여행(tour) 과정을 움직이는 그림으로 보여준다 (파일: biplot oliveoil local pca.avi).

References

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association*, **80**, 580–598.

Cook, D. and Swayne, D. F. (2007). *Interactive and Dynamic Graphics for Data Analysis*, Springer.

Huh, M. H. and Lee, Y. (2012). Local projective display of multivariate numerical data, *Korean Journal of Applied Statistics*, **25**, 661–668. (Written in Korean).

Hurley, C. B. (2004). Clustering visualization of multidimensional data, *Journal of Computational and Graphical Statistics*, **13**, 788–806.

Wu, H. M., Tzeng, S. L. and Chen, C. H. (2008). Matrix visualization, In *Handbook of Data Visualization* (edited by C. H. Chen, W. K. Hardle, and A. Unwin), Springer, 681–708.

움직이는 데이터 그림

허명회[a,1]

[a]고려대학교 통계학과

(2013년 10월 8일 접수, 2013년 11월 16일 수정, 2013년 11월 21일 채택)

요 약

이 연구는 다음 몇 가지 경우에 적용 가능한 '움직이는 데이터 그림(moving data pictures)'을 제안 한다: 1) 한국어 텍스트의 단어 구름(word cloud), 2) $n \times p$ 행렬의 시각화(matrix visualization), 3) $p \times p$ 산점도 행렬의 동영상 버전, 4) k개 개체 군집의 동적 시각화 등. 이들 기법은 데이터에 내재된 숨은 정보와 시각적 아름다움을 드러내고 정보 소비자들의 흥미를 점화할 수 있다.

주요용어: 자료 시각화, 동적 그래픽스, 움직이는 그림, 단어구름, 열지도, 산점도행렬, k-평균 군집화.

[1](136-701) 서울시 성북구 안암동 5가 1, 고려대학교 정경대학, 교수. E-mail: stat420@korea.ac.kr