The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.4.661
25(4), 661–668

Local Projective Display of Multivariate Numerical Data

Myung-Hoe Huh[1] · Yonggoo Lee[2]

[1]Department of Statistics, Korea University; [2]Department of Statistics, ChungAng University

(Received May 25, 2012; Revised July 10, 2012; Accepted July 23, 2012)

Abstract

For displaying multivariate numerical data on a 2D plane by the projection, principal components biplot and the GGobi are two main tools of data visualization. The biplot is very useful for capturing the global shape of the dataset, by representing n observations and p variables simultaneously on a single graph. The GGobi shows a dynamic movie of the images of n observations projected onto a sequence of unit vectors floating on the p-dimensional sphere. Even though these two methods are certainly very valuable, there are drawbacks. The biplot is too condensed to describe the detailed parts of the data, and the GGobi is too burdensome for ordinary data analyses. In this paper, "the local projective display(LPD)" is proposed for visualizing multivariate numerical data. Main steps of the LDP are 1) k-means clustering of the data into k subsets, 2) drawing k principal components biplots of individual subsets, and 3) sequencing k plots by Hurley's (2004) endlink algorithm for cognitive continuity.

Keywords: Data visualization, biplot, GGobi, supplementary data, endlink algorithm.

1. 연구배경과 목적

n개 관측 · p개 변수의 다변량 수치 자료 X에 대한 차원 주성분 행렬도(principal components biplot)는 다음과 같이 구성된다 (Gabriel, 1971; Choi, 1999; Huh, 2011).

$$\text{행(관측) 점: } XV^{(s)}\text{의 } n\text{개 행}, \qquad \text{열(변수) 점: } V^{(s)}\text{의 } p\text{개 행},$$

여기서 자료 행렬 X의 p개 열(변수)이 표준화되어 있음이 가정되었고 $V^{(s)}$는 X^tX의 계수(rank) s 주고유벡터 행렬이다. 실용적 편의성의 이유로, 2차원 이외의 주성분 행렬도가 고려되는 경우는 드물다. 이 연구에서도 $s = 2$를 가정한다. 열 점은 변수의 방향과 척도를 나타내므로 상수배로 늘여도 된다. 이 연구에서는 행 점들과의 겹침을 피하기 위해서 3.5배로 연장된 곳에 열 점을 찍는다. n_+개 관측의 추가 자료(supplementary data) X_+는 동일 플롯에 $X_+V^{(s)}$의 p개 행을 타점하여 표출할 수 있다.

G고비(GGobi)는 X고비(XGobi)의 진화 버전으로 다변량 자료의 2차원 동적 사영(dynamic projection)을 구현한 오픈소스의 자료 시각화 소프트웨어이다 (Buja 등, 1998; Cook과 Swayne, 2007). G고비는 R의 `rggobi` 팩키지와 연동하여 p차원 구(球) 상에서 사영벡터의 임의적 움직임에 따른 n개 관측

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

점의 사영을 실시간으로 보여준다 (http://www.ggobi.org). Holes 지수 또는 LDA 지수 등 일정한 기준에 따라 최적 사영을 추구할 수도 있다.

본 연구는 실제 응용에서 행렬도가 자료의 세부적 모습을 보기에는 너무 압축적이고 G고비는 부담이 크다는 관점에서 출발한다. 주성분 행렬도는 다변량 자료에 내재된 주성분이 2개로 충분한 경우가 아니면 전체자료를 표출해내는 데 있어 한계가 있다. G고비는 매우 많은 수의 사영도를 짧은 시간 간격으로 보여주기 때문에 자료 분석자가 인지적으로 소화해내기 어렵다.

제안 방법 '국소적 사영 전시'(local projective display)는 전체 자료를 k개로 군락화하고 군락 자료를 개별적으로 주성분 행렬도로 표출하면서 잔여 자료를 동일 플롯의 배경에 넣는다. 또한 Hurley의 끝잇기(endlink) 알고리즘으로 k개 그래프를 순서화함으로써 정보흐름의 인지적 연속성을 추구한다. 이에 따라 전체 자료의 국소적(local) 모습을 포착할 수 있고 각 부분이 전체 중 어디에 있는가를 알 수 있다.

2절에서 기본 방법론을 제안하고 수치 예를 제시한다. 3절에서 방법론의 일부를 보완하고 4절과 5절에서 모의생성 자료 사례와 실제 자료 사례에 제안 방법론을 적용해 보인다.

2. 제안 방법론: 기본

실제의 다변량 수치 자료는 다수 군락(cluster)의 집합으로, 각 군락은 각각 다른 구조로 되어 있을 수 있다. 특히 자료 탐색(data exploration)의 단계에서는 그런 관점이 요구된다. 본 연구자는 다음과 같이 개별 군락 중심적인 자료 시각화 방법을 제안한다.

1) 전체 자료를 k개 군락 $C_1, \ldots, C_k$로 분할하고 군락 중심(centroid) $\mathbf{m}_1, \ldots, \mathbf{m}_k$를 구한다. 군락 C_j의 크기를 n_j로 표기한다.
2) 크기 n_j의 副표본(subsample) S_j를 각 $\mathbf{m}_j$를 중심으로 구성한다 ($j = 1, \ldots, k$). 부표본 S_j의 자료를 열 중심화하여 $n_j \times p$ 행렬 X_j로 표기한다.
3) 각 $j\,(= 1, \ldots, k)$에 대하여, $X_j^t X_j$를 고유 분해하여 $V_j^{(s)}$를 산출하고 전체 자료 X를 $V_j^{(s)}$의 각 열에 사영한다. 이에 따라 s차원 평면에

$$\text{행(관측): } XV_j^{(s)}\text{의 } n\text{개 행}, \qquad \text{열(변수): } V_j^{(s)}\text{의 } p\text{개 행}$$

이 타점된다. 이 때, 군락 C_j의 관측은 前面 이미지로, 그 외 관측은 後面 이미지(배경)로 대비되도록 한다.

단계 2에서 부표본 S_j를 군락 C_j와 일치시키지 않고 군락 중심 $\mathbf{m}_j$로부터 가장 가까운, 즉, 유클리드 거리가 가장 작은, n_j개의 관측들로 구성하는 이유는 $S_1, \ldots, S_k$를 $C_1, \ldots, C_k$와 일치시키는 경우 2개 이상의 군락이 접하는 상황에서는 군락의 공분산 구조가 왜곡되기 때문이다. 그렇지만 부표본 S_j의 크기는 군락 C_j의 크기와 일치시켰는데 그것이 부표본 S_j의 크기를 '대체로' 나타내는 데는 문제가 없을 것으로 생각된다.

이 연구는 자료 시각화에 초점을 맞출 것이므로 군락의 수 k의 선택에 대한 방법론과 논의를 하지 않으려 한다. 자료 시각화의 결과가 군락의 수 k를 변경할 필요가 있다는 것일 수 있지만, 이 문제는 또 하나의 큰 이슈이기 때문이다.

수치 예: iris 자료

피셔의 붓꽃(iris) 자료는 3개 종(species; setosa, versicolor, virginica)의 150개 개체에 대한 4개 수치

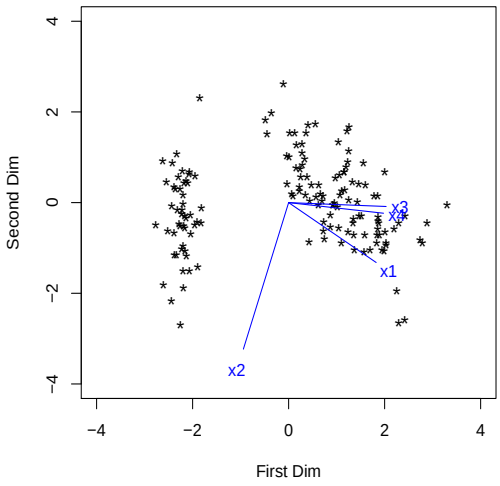

Figure 2.1. Principal components biplot of the iris data

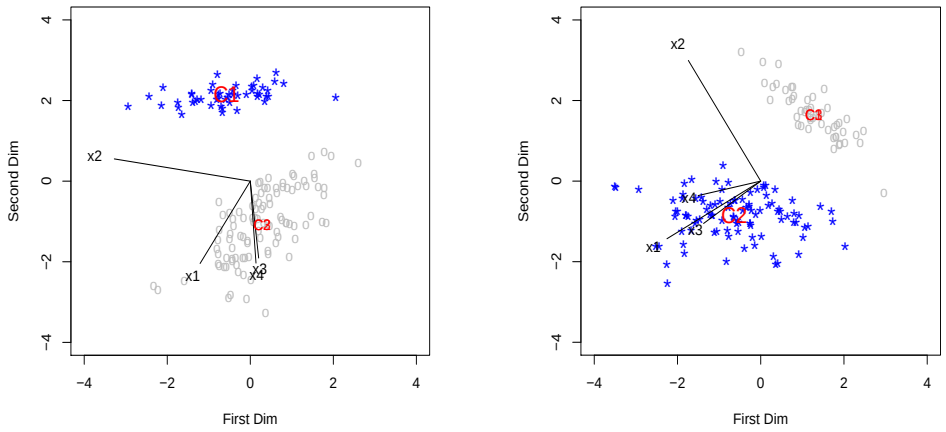

Figure 2.2. Local projective display of the iris data, $k = 2$

변수로 구성되어 있다 (x1 = sepal length, x2 = sepal width, x3 = petal length, x4 = petal width). 여기서는 종(種) 변수를 사용하지 않는다.

Figure 2.1은 전체 자료에 대한 주성분 행렬도이다. 최소한 2개 군락이 있음이 확연하다. 그림 왼쪽의 작은 군락은 오른쪽의 큰 군락에 비하여 x1, x3, x4의 값이 상대적으로 작다. x2의 값에서는 왼쪽 군락이 오른쪽 군락에 비해 약간 크다.

Figure 2.2는 iris 자료에 대한 $k = 2$인 국소적 사영 전시이다. 왼쪽 그래프에서는 군락 C1에 포커스가 두어졌고 오른쪽 그래프에서는 군락 C2에 포커스가 두어졌다. 군락 중심점에는 군락 레이블이 찍혔다.

두 그래프가 사영 방향에서 다소 차이가 있지만 결과적인 모습은 크게 다르지 않다. 왼쪽 그래프를 시계 바늘 회전 방향으로 약 45도 회전하면 오른쪽 그래프가 된다.

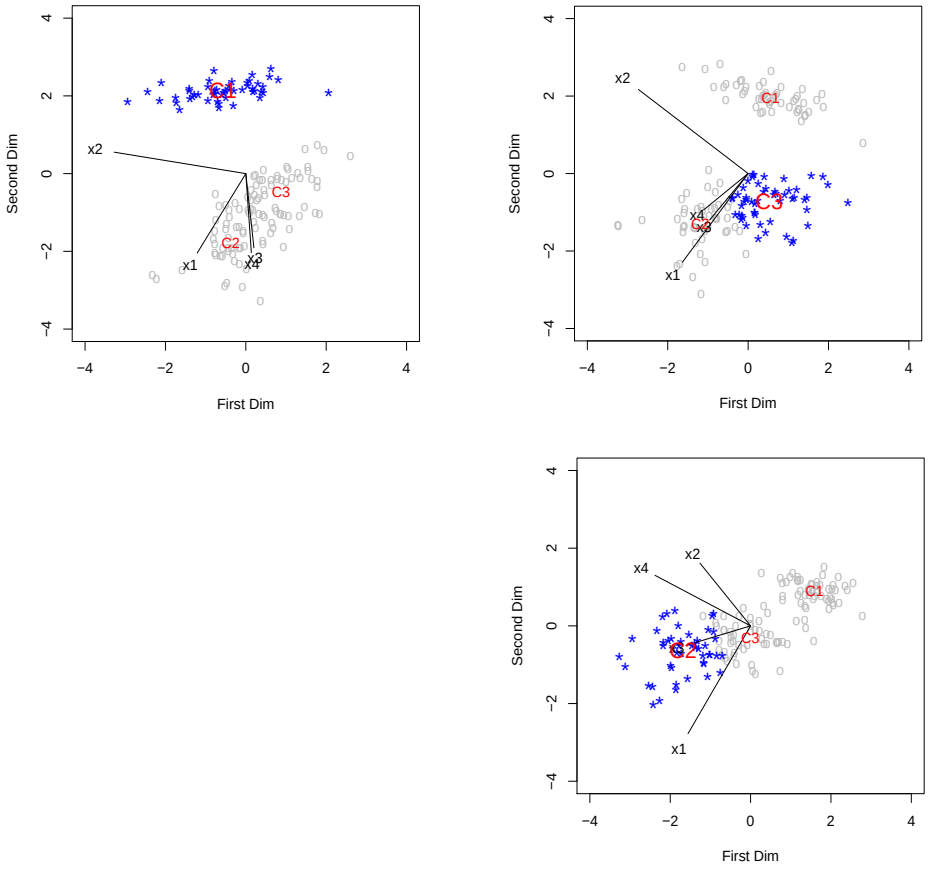

Figure 3.1. Local projective display of the iris data, $k = 3$ (Left upper → Right upper → Right lower)

3. 제안 방법론: 보완

앞 절의 기본 방법론을 두 가지 점에서 보완한다. 자료분석자의 인지적 해석이 편리하도록, k개 그래프의 전시 순서를 정하고 그래프의 상하좌우를 조정하자는 것이다.

1) 기본 방법론에서 k-평균 군락화를 쓰므로 군락 레이블은 임의적(arbitrary)으로 정해진다. 인지적 연속성을 위하여 k개 그래프의 순서를 재배열할 필요가 있다.

   Hurley (2004)의 끝잇기(endlink) 알고리즘은 유사성 기준에 기반하여 총 유사성이 가장 크게 되도록 k개 오브젝트를 일렬로 배열해준다. 따라서 음부호의 중심 간 거리를 군락 간 유사성으로 정의하여 k개 군락에 끝잇기 알고리즘을 적용한다. 그러면 인접 군락 간 거리(음의 유사성)의 합이 가장 작은 군락 레이블의 순열(permutation)이 산출된다. 끝잇기(endlink) 알고리즘은 R의 `gclus` 팩키지에 구현되어 있다 (http://www.r-project.org).

2) 기본 방법론의 단계 3에서 고유벡터 행렬 $V_j^{(s)}$의 열 부호는 임의적(arbitrary)이다. 즉, $V_j^{(s)}$를 $V_j^{(s)} D_{\pm 1}^{(s)}$로 대체해도 마찬가지이다. 여기서 $D_{\pm 1}^{(s)}$은 대각요소가 1 또는 -1인 $s \times s$ 대각행렬이다. 이 점에 착안하여, 연이은 2개 그래프의 제1축 사영 간 상관이 플러스 부호를 취하도록 $V_j^{(s)}$의 제1열을

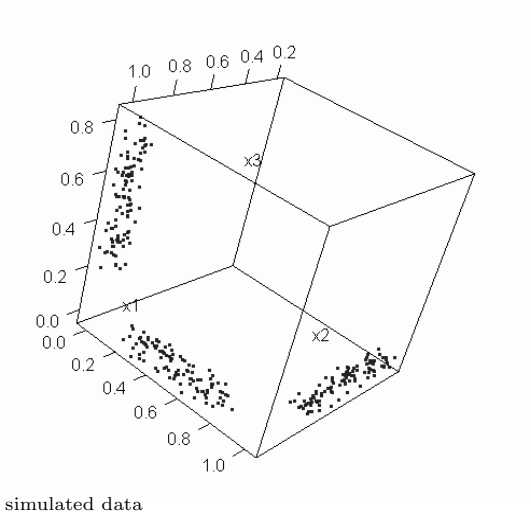

Figure 4.1. 3D plot of the simulated data

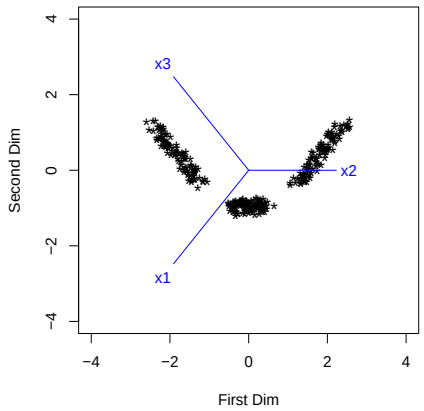

Figure 4.2. Principal components biplot of the simulated data

부호화함으로써 인지적 연속성을 추구한다. 마찬가지 방법으로 연이은 2개 그래프의 제2축 사영 간 상관의 부호도 플러스가 되도록 한다.

Figure 3.1은 iris 자료에 대한 $k = 3$인 국소적 사영 전시이다. 끝잇기 알고리즘을 적용한 결과 군락순서가 C1–C3–C2로 나타나서 이 순서로 사영도가 배열되어 있다 (Left upper → Right upper → Right lower). 군락 C3의 한쪽 경계와 군락 C2의 한쪽 경계가 겹쳐 있으나 3개 변수 x1, x3, x4의 값에서 C2가 C3에 비해 다소 큰 경향이 있음이 발견된다. 변수 x2에서는 C2와 C3가 별로 다르지 않다.

4. 모의생성 자료 사례

이 절에서는 모의생성 자료에 기존의 주성분 행렬도와 제안 방법론을 적용하여 볼 것이다. 모의생성 자료는 300개의 3차원 관측인데, Figure 4.1에서와 같이 변 길이 1의 정육면체에서 4개의 꼭지점 $(0, 1, 0)$, $(1, 1, 0)$, $(1, 0, 0)$, $(1, 0, 1)$을 잇는 3개 변에서 mid point 중심의 ± 0.3의 균일 임의수와 0 중심

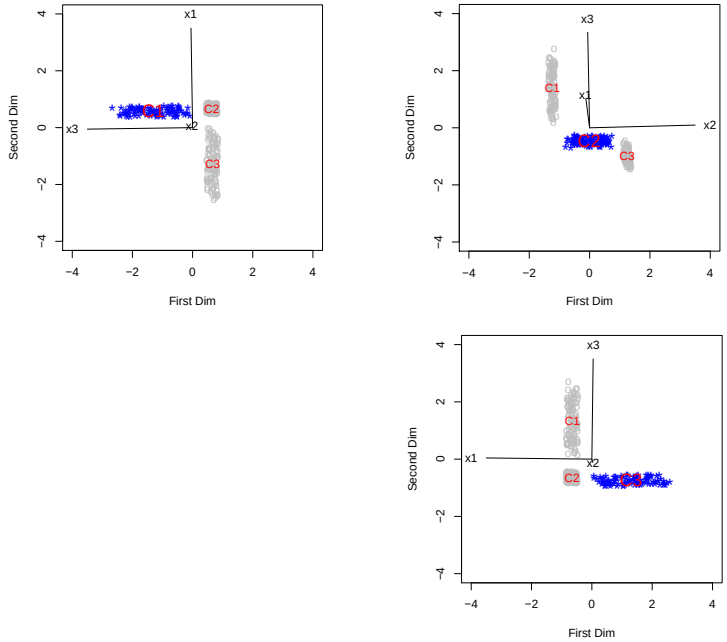

Figure 4.3. Local projective display of the simulated data, $k = 3$ (Left upper → Right upper → Right lower)

의 ±0.05의 균일 임의오차를 붙이는 방식으로 생성되었다.

Figure 4.2는 이 자료에 대한 주성분 행렬도인데 그럴 듯해 보이기는 하지만 자료분석자에게 그릇된 정보를 준다. 왜냐하면 이 모의생성 자료는 2차원 사영으로는 주 특성이 포착될 수 없기 때문이다.

Figure 4.3은 이 자료에 제안 방법론을 적용한 결과이다. 군락 C1이 변수 x3를 따라, 군락 C2는 변수 x2를 따라, 군락 C3는 변수 x1을 따라 산재되어 있음을 볼 수 있다.

5. 이탈리아 올리브오일 자료 사례

이탈리아 남부 4개 지역(Calabria, North Apulia, Sicily, South Apulia)에서 생산된 323종 올리브油의 8개 지방산 성분(fatty acid composition) 자료에 제안 방법론을 적용해보기로 한다 (Cook과 Swayne, 2007; x1 = palmitic, x2 = palmitoleic, x3 = stearic, x4 = oleic, x5 = linoleic, x6 = linolenic, x7 = arachidic, x8 = eicosenoic). 군락 수 k는 4로 하였다. Figure 5.1이 군락자료 각각에 대한 국소적 사영으로 얻은 결과이다 (Left upper → Right upper → Right lower → Left lower).

각 군락이 구분되어 나타났고 각 군락의 특성을 살펴볼 수 있다. 예컨대 군락 C3는 x4, x6, x7, x8이 크다는 점에서 군락 C1, C2, C4와 다르다. Figure 5.2는 전체자료에 대한 주성분 행렬도인데 4개 군락들이 상당히 겹쳐져 있어 군락들의 개별적 특성을 추려내기 쉽지 않다.

6. 맺음말

이 연구에서 제안된 국소적 사영 전시는 k개 군락 각각에 한 차례씩 포커스를 두어 자료의 국소적 특성을 살펴보는 그래픽스 기법이다. 또한 일련의 그래프에서 군락들의 상호적 위치 관계를 파악할 수 있다.

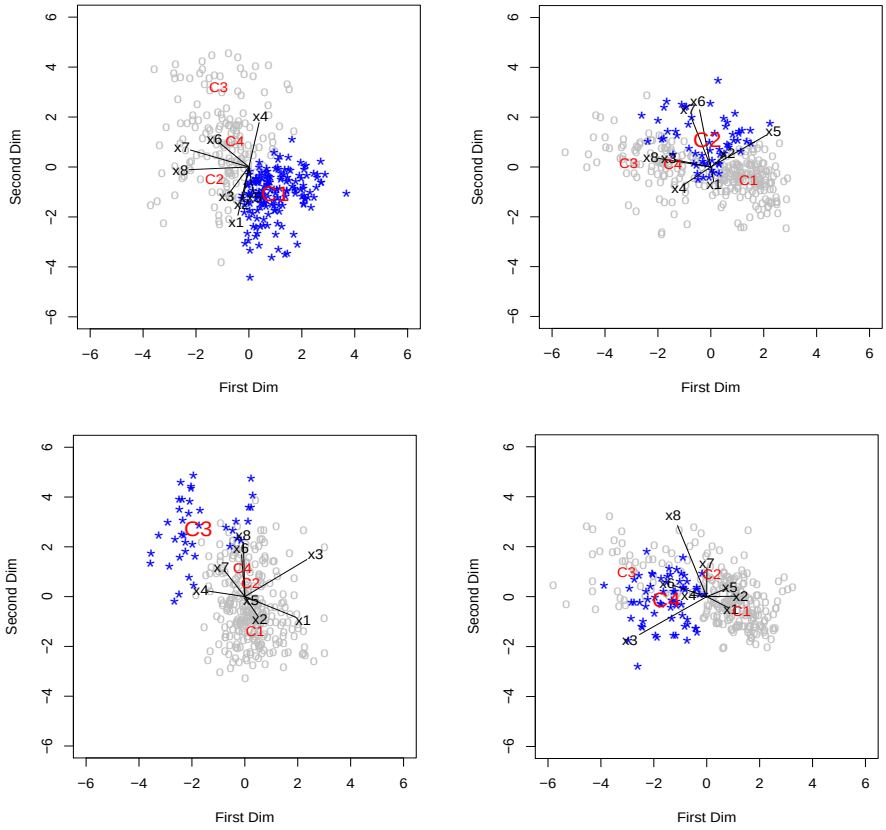

Figure 5.1. Local projective display of the Italian olive oil data, $k = 4$ (Left upper → Right upper → Right lower → Left lower)

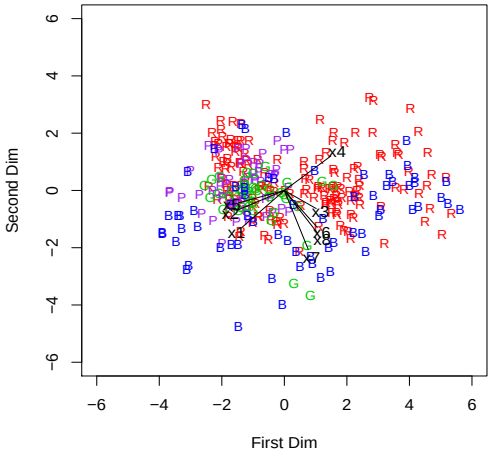

Figure 5.2. Principal components biplot of the Italian olive oil data (C1 "red" r, C2 "purple" p, C3 "green3" g, C4 "blue" b)

따라서 전체 자료를 1개 그래프로 시각화하는 주성분 행렬도에 비교하여 보다 풍부한 자료해석이 가능하다.

이론적으로는 국소적 사영 전시를 동적 그래픽스 기법인 G고비(GGobi)의 수많은 화면 중 k개의 특수한 화면으로 볼 수 있겠지만 실제로는 수작업만으로 그와 같은 화면을 만들기 어려울 것이다.

References

Buja, A., Cook, D. and Swayne, D. F. (1998). XGobi: Interactive Dynamic Data Visualization in the X Window System, *Journal of Computational and Graphical Statistics*, **7**, 113–130.

Choi, Y. S. (1999). *Understanding Biplots and Applications* (written in Korean), Busan National University Press.

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

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58**, 453–467.

Huh, M. H. (2011). *Exploratory Multivariate Data Analysis* (in Korean), Freedom Academy, Korea.

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