응용통계연구 제 9권 1호
1996년 3월 pp. 45-52.

동적 평행좌표그림과 그의 활용

장 대 홍[1], 양 수 정[2]

요 약

자료의 구조와 특징을 파악하기 위한 탐색적 자료분석을 행할 때 유용한 수단으로 통계 그래픽스가이용된다. 평행좌표그림(parallel coordinate plot)을 통계 그래픽스의 한 방법인 동적 그래픽스로서 이용하기 위하여 동적 평행좌표그림이 쓰이는데, 이 도구를 이용하면 3차원 이상의 다차원 자료를 동적 그래픽스로 표현, 분석할 수 있다. 본 논문에서는 하나의 동적 평행좌표그림을 제시하고, 자료분석의 예를 보였다.

1. 서론

우리는 자료들의 구조를 파악하기 위해서나 그 자료들로부터 다양한 정보를 알아보기 위하여 여러가지의 통계적 도구들을 이용한다. 자료를 분석함에 있어서는 흔히, 자료의 구조와 특징을 파악하기 위한 탐색적 자료분석의 단계와, 관찰된 형태나 효과의 재현성을 평가하는 확증적 자료분석의 단계로 구분할 수 있다. 탐색적 자료분석에서는 우리가 분석하고자 하는 자료들의 구조를 찾는 것이 제일 중요한 문제인데, 이러한 탐색적 자료분석시 유용한 도구가 통계 그래픽스이다. 통계그래픽스는 통계적 정보전달의 강력한 도구이며, 변수들 사이의 관계를 나타내고자 하거나, 자료를 분석하거나,분석한 자료의 결과를 전달하거나 전시하는데 사용된다. 그 예로 산점도(scatter plot)를 들 수 있다. 산점도는 자료를 표현하기에 쉬우므로, 널리 사용되는 자료표현의 도구이며, 또한 산점도로 표현된 자료들은 분석하기에도 용이하다. 이런 산점도는 데이타의 관찰이 용이하다는 장점을 가진 반면 2차원 이상의 자료를 표현하기가 어렵다는 단점을 지니고 있다. 따라서 삼변량이상의 다변량의 자료들을 처리할 수 있는 통계 그래픽스 도구를 필요로 하게 된다. 다변량의 자료들을 처리하는 그림도구로서는 대표적으로 Chernoff 얼굴(1973), glyphs, 별그림(star diagrams), Andrews 곡선(1972)등 여러가지가 있다. 또한 동적 그래픽스(dynamic graphics)의 방법으로 회전(rotation)과 붓질(brushing)기법(Becker와 Cleveland(1988)), grand tour(Buja, Asimov, Hurley와 McDonald(1988)), R-code(Cook과 Weisberg (1989))등이 있다. 다변량의 자료들을 처리하기 위한 방법으로 제시한 회전은 3차원의 직각좌표계상에 자료들을 표시하고, 각 축및 임의의 가상축에 대해서 회전시키는 방법이다. 산점도가 2차원에 국한되는 점을 보완한 도구 중 산점도 행렬이 있는데, 이는 다차원의 자료표현이 가능하다. 또한, 산점도 행렬에서 3변수 혹은 4변수이상의 자료들의 구조를 한꺼번에 알기

1) (608-737) 부산광역시 남구 대연3동 599-1, 부산수산대학교 자연과학대학 응용수학과 부교수.
2) 부산수산대학교 자연과학대학 응용수학과 석사과정 졸업.

위하여 붓질기법이 제안되었다(Becker와 Cleveland(1987)).

우리는 또 다른 동적 그래픽스의 방법으로서 동적 평행좌표그림(dynamic parallel coordinate plot)을 제안할 수 있다. 이 도구는 다차원의 자료를 표현하는 데 있어서 효과적인 도구이다. Inselberg(1985)의 논문에서 최초로 이런 평행좌표그림을 언급한 바가 있다. 또한, Wegman(1990)의 논문에서는 평행좌표그림이 통계분석을 함에 있어서 유용한 도구로써 활용되며, 특히 군집분석을 행할 때 이용하면 효과적이라고 서술하고 그림그리기(painting) 기능이라는 동적 그래픽스 기능을 이용하여 하나의 동적 좌표그림을 제시하고 있고, Gennings, Dawson, Carter와 Myers (1990)의 논문에서는 평행좌표 그림을 생물통계 분석에 이용하였다. 본 논문에서 통계 그래픽스의 한 방법인 동적 그래픽스로서 이용하기 위하여 하나의 동적 평행좌표그림을 제안하고, 여러 통계자료분석에 동적 평행좌표그림을 활용하는 방법을 예와 더불어 제시하였다.

2. 동적 평행좌표그림

2.1 평행좌표그림의 개념 및 성질

우리는 유클리드 공간상의 직각좌표계를 일반적으로 다루어 왔다. 그러한 2차원 XY-평면상에서 Y축에 평행하게 X축의 양의 방향으로 등간격으로 N개의 선을 그어서 각각 $X_1, X_2, \cdots, X_N$을 대응시킨다. 그러면 N차원의 새로운 좌표축 $X_1, X_2, \cdots, X_N$이 생기게 되는데, 이 좌표축을 평행좌표축이라 한다. N차원 유클리드 공간상의 한 좌표 $C: (c_1, c_2, \cdots, c_N)$은 평행좌표축에 그림1과 같이 대응되게 된다. 즉, 유클리드 공간의 좌표 C의 한 좌표값 c_1은 평행좌표계상의 X_1축상의 좌표 $(0, c_1)$에 대응되고, c_2는 X_2축상의 $(1, c_2)$에, c_i는 X_i축상의 $(i-1, c_i)$에 각각 대응된다. 이 때, N차원의 유클리드 공간상의 한 좌표 C는 평행좌표계에서는 그림 1처럼 대응된 각 좌표를 연결하여 만드는 꺾은선(polygonal line)으로 재표현되는 것이다. 즉, 유클리드 공간상의 한 점 C는 평행좌표축에서는 한 꺾은선에 대응되는 것이다.

평행좌표계는 다음과 같은 쌍대성 성질들을 갖는다(Inselberg(1985)).

① 점과 선의 쌍대성
② 이동과 회전의 쌍대성
③ 원추곡선들 사이의 쌍대성
④ 첨점과 변곡점의 쌍대성

첨점과 변곡점의 쌍대성을 예로 들면, 일변량 표준정규분포에서는 직각좌표계상에서의 X_1축의 값이 -1과 1인 지점에서 각각 하나씩 두개의 변곡점이 나타난다. 그림 2는 표준정규분포를

평행좌표계로 나타낸 평행좌표그림이다. 직각좌표계상의 두 개의 변곡점이 평행좌표계상의 두 개의 첨점으로 대응됨을 확인하여 볼 수 있다.

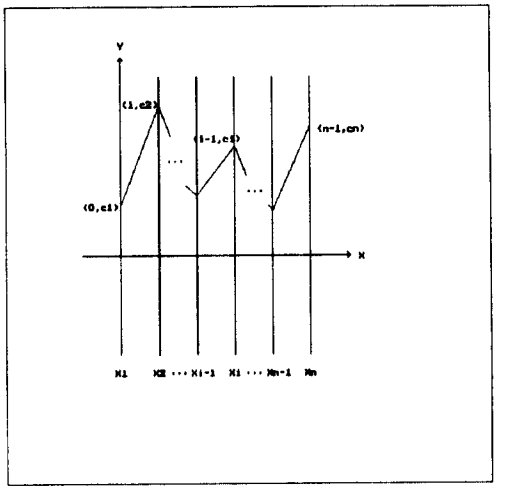

그림 1. N차원의 평행좌표계

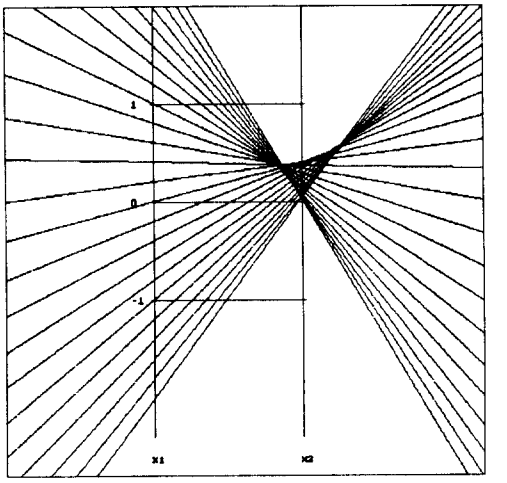

그림 2. 평행좌표계상에서의 표준정규분포

2.2 동적 평행좌표그림

평행좌표계의 개념을 기초로 동적 평행좌표그림(dynamic parallel coordinate plots,약자로 DPCP.) 프로그램을 작성하였는데, 이 DPCP프로그램은 Becker와 Cleveland(1987)가 제안한 붓질기법에서 나오는 네가지의 붓질작업(brushing operations)(밝게하기(highlight), 그림자 밝게하기(shadow highlight), 생략하기(delete), 이름붙이기(label))과 세가지의 그리기방법(paint modes) (일시(transient), 연속(lasting), 취소(undo))의 개념을 기초로 했으며, TURBO C를 사용하여 짠 프로그램이다. 이런 기법들은 컴퓨터 입력장치 중의 하나인 마우스(mouse)의 이용을 전제로 한 프로그램이다. 물론, 여기서의 세가지 그리기방법들은 자료분석가의 필요에 따라 선택한 각각의 작업 주메뉴에 공통적으로 택할 수 있는 부메뉴들로서, 선택한 후 자료의 변화를 관찰, 검토하여 분석하게 된다.

본 논문에서 DPCP 프로그램을 다루는 이유는 다변량의 자료들의 표현이 쉽고, 사용자가 필요한 만큼 원하는 축들을 선택하여 조건을 준 다음, 선택되지 않은 축들사이의 자료들의 관계를 판단할 수 있는 장점들이 있기 때문이다.

그림 3은 DPCP 프로그램의 메뉴 중 밝게하기 붓질작업을 선택하고 일시 그리기방법을 이용하여 Cook과 Weisberg(1982)의 책에 나오는 전투기 자료에서 6번째 변수값이 1인 자료, 즉 항공모함에 착륙할 수 있는 비행기들을 표시한 화면을 나타내고 있다.

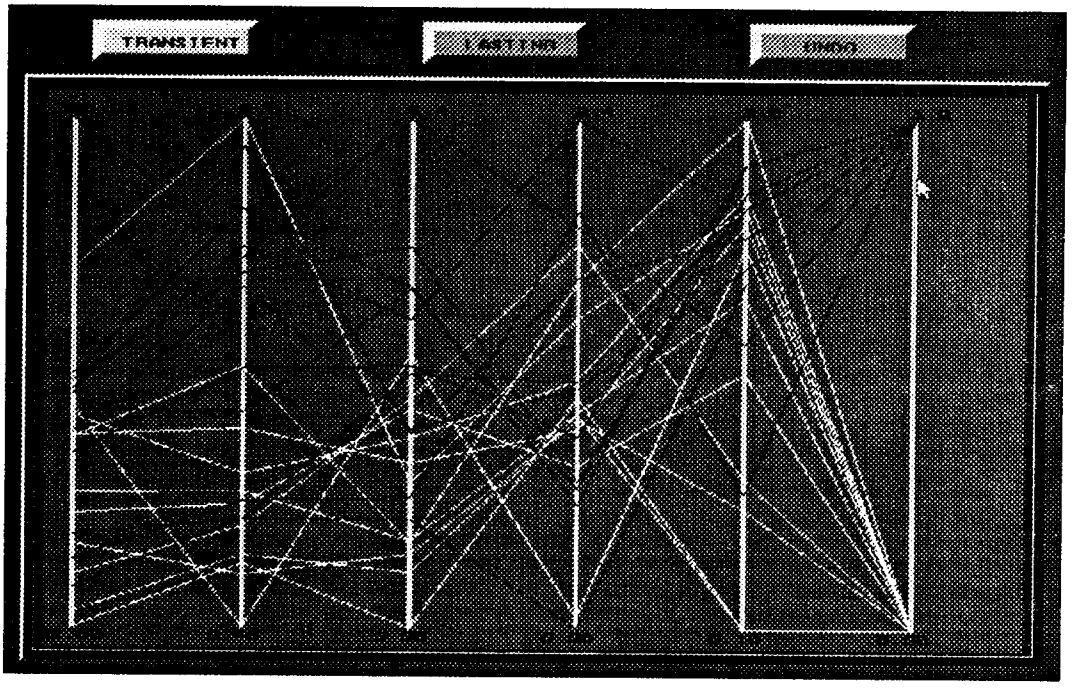

그림 3. DPCP프로그램에서 밝게하기 붓질작업을 실행시켰을 때의 화면

3. 동적 평행좌표그림의 통계에의 응용

동적 평행좌표그림은 다변량 자료를 표현하는데 효과적인 그림도구이므로 다변량 자료의 탐색을 위한 동적 그래픽스 방법으로 이용할 수 있고, 다변량 통계분석의 보조 수단으로 쓸 수 있다. 본 논문에서는 중회귀분석과 혼합물자료분석(compositional data analysis)에서 동적 평행좌표그림을 이용한 예들을 들고자 한다.

중회귀분석에서 설명변수들과 반응변수 사이의 선형관계를 알아보거나 조건부 분포를 알고자 할 때 동적 평행좌표그림을 이용하면 편리하다.

예 1. 다음의 예는 Draper와 Smith(1981) 책에 나오는 예로서, 설명변수가 4개이고, 반응변수가 1개인 회귀분석모형을 위한 자료이다. 이 자료를 DPCP로 표현한 것이 그림 4이다. 첫번째 축이 설명변수 x_1으로서 월별 평균온도, 두번째 축이 설명변수 x_2로서 생산량(M pounds), 세번째 축이 설명변수 x_3로서 월별 공장운행일수, 네번째 축이 설명변수 x_4로서 월별 종사자

수이다. 다섯번째 축이 반응변수 y로서 월별 물사용량(gallons)이다. 이 PCP에서 생략하기 붓질작업을 선택하고 연속 그리기방법를 이용해서 x_1값이 중간이하 부분, x_2값이 중간부분, x_3값이 중간이하 부분이라는 3가지 조건하에 x_4와 y와의 관계를 나타낸 DPCP가 그림 5이다. 이 그림에서 알 수 있는 것은 3가지 조건하에서 x_4와 y사이에 전반적으로 선형관계가 성립된다는 사실이다.

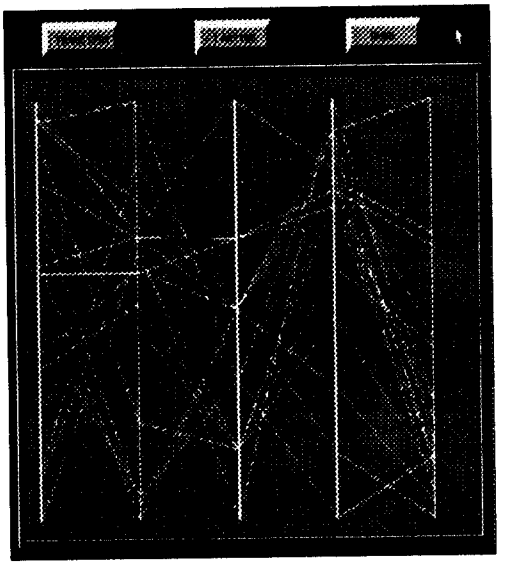

그림 4. 예 3.1에 대한 초기화면

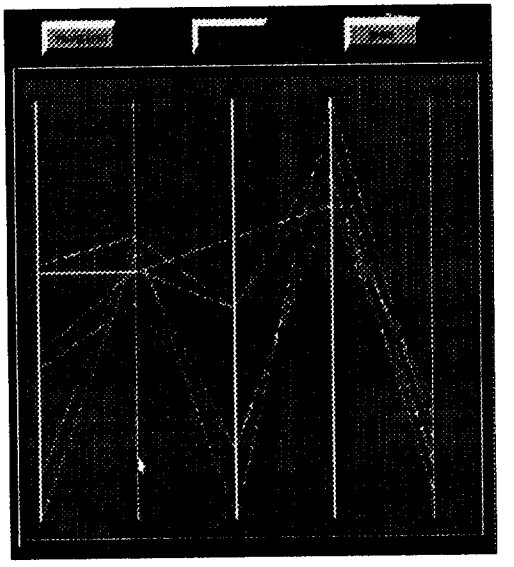

그림 5. 3가지 조건을 부여했을때의 DPCP화면

혼합물자료분석에서는 임의의 벡터가 비율을 나타내는 비음요소인 각 구성성분 $x_1, x_2, \cdots, x_D$로 구성되어 있는 자료들이다. 즉, 다음과 같은 제약 조건이 성립한다.

$$x_1 + x_2 + \cdots + x_D = 1$$

이런 자료에 대한 통계적 분석에 들어가기 전에 탐색적 단계로서 그림으로 형상화시킬때 $D \geq 4$이면 그리기가 어렵고, $D \geq 5$이면 전체 자료를 한 그림에 그릴 수 없게 된다. 통계적 분석시에도 D가 클 때에는 주로, 부구성(subcomposition)과 분할(partition)을 이용하여 구성성분의 수를 줄여 분석을 행한다. 이 때, 동적 평행좌표그림을 이용하면 탐색적 단계에서 유용하다.

예 2. 그림 6과 그림 7은 Aitchison(1986) 책에 나타나는 자료로서, 남극의 호수에서 3가지

구성물질(모래, 침니, 진흙)로 구성된 39개의 침전물을 물깊이에 따라 나타낸 DPCP들이다. 이 그림들에서 밝게하기 붓질작업을 선택하고 일시 그리기방법를 이용하면 모래와 침니는 서로 반비례적인 관계를 갖고 있음을 알 수 있고, 물깊이의 변화에 따른 구성비의 변화도 알 수 있다.

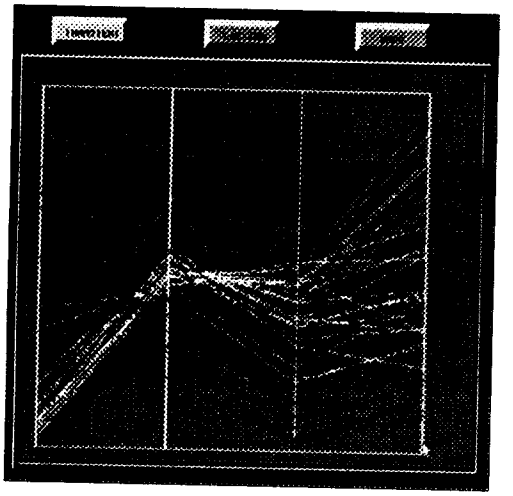

그림 6. 물깊이가 낮은 경우의 DPCP화면

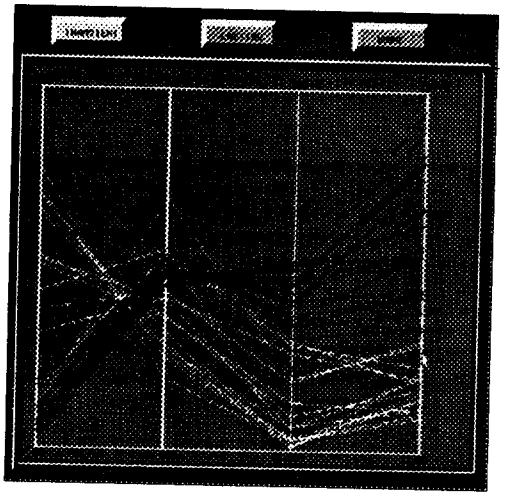

그림 7. 물깊이가 높은 경우의 DPCP화면

4. 결 론

우리가 표현, 분석하고자하는 자료가 다변량인 경우, 통계 그래픽스에서는 지금까지 여러가지 동적 그래픽스 방법들이나 다차원을 표현할수 있는 도구들을 서로 보완하면서 이용하여 왔다. 우리는 동적 평행좌표그림을 다변량 자료의 탐색을 위한, 하나의 동적 그래픽스 방법으로서 이용할 수 있다. 단지, 우리가 교육받아와서 익숙해져 있는 직각좌표계의 관점에서 벗어나서, 조금 다른 시각으로 평행좌표계의 기본 성질들을 이용하면 표현된 자료들의 상태나 의미 등을 해석할 수 있게 된다.

참고문헌

[1] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*, Chapman and Hall, London.

[2] Andrews, D. F. (1972). Plots of high-dimensional data, *Biometrics*, Vol. 28, 125-136.

[3] Becker, R. A., Cleveland, W. S. (1987). Brushing Scatterplots, *Technometrics*, Vol. 29, 127-142.

[4] Becker, R. A., Cleveland, W. S. (1988). The Use of Brushing and Rotation in Data Analysis, *Dynamic Graphics for Statistics*, eds. W. S. Cleveland and M. E. McGill, Wadsworth/Brooks Cole, Monterey, CA, 247-275.

[5] Buja, A., Asimov, D., Hurley, C., and McDonald, J. A. (1988). Elements of a Viewing Pipeline for Data Analysis, *Dynamic Graphics for Statistics*, eds. W. S. Clevelend and M. E. McGill, Wadsworth/Brooks Cole, Monterey, CA, 277-308.

[6] Chernoff, H. (1973). Using faces to represent points in K-dimensional space graphically, *Journal of American Statistical Association*, Vol. 68, 361-368.

[7] Cleveland, W. S. and McGill, M. E.(eds.) (1988). *Dynamic Graphics for Statistics*, Wadsworth/Brooks Cole, Monterey.

[8] Cook, R. D. and Weisberg, S. (1982). *Residuals and Inflence in Regression*, Chapman and Hall, London.

[9] Cook, R. D. and Weisberg, S. (1989). Regression Diagnostics with Dynamic Graphics (with discussion), *Technometrics*, Vol. 31, 277-310.

[10] Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, 2nd ed, John Wiley and Sons, New York.

[11] Gennings, C., Dawson, K. S.,Carter, W. H. and Myers, R. H. (1990). Interpreting Plots of a Multidimensional Dose-Response Surface in a Parallel Coordinate System, *Biometrics*, Vol. 46, 719-735.

[12] Inselberg, A. (1985). The Plane with Parallel Coordinates, *The Visual Computer*, Vol. 1, 69-91.

[13] Wegman, E.J. (1990). Hyperdimensional Data Analysis Using Parallel Coordinates, *Journal of the American Statistical Association*, Vol. 85, 664-675.

The Dynamic Parallel Coordinate Plot and Its Applications

Dae-Heung Jang[1], Soo-Jeong Yang[2]

Abstract

In this paper, we describe the basic properties of the parallel coordinate plots and propose a dynamic parallel coordinate plot. This dynamic parallel coordinate plot can be used as a dynamic graphics tool for multivariate data analysis.

1) Associate Professor, Dept. of Applied Mathematics, College of Natural Science, National Fisheries University of Pusan, Pusan, 608-737.

2) Master, Dept, of Applied Mathematics, College of Natural Science, National Fisheries University of Pusan.