한국통계학회논문집 제2권 2호
1995년 12월 pp. 423-433

다차원 자료의 구조탐색에서 통계 그래픽스 방법의 활용1)

허문열, 이경미2)

요 약

자료분석에서 다루는 차원의 수는 무한히 클 수 있다. 따라서 컴퓨터 그래픽스 분야에서 많이 연구하고 있는 결과를 그대로 적용하는 것은 한계가 있다. 이에 통계학자들은 다차원 자료의 구조 탐색을 위해 여러 가지 간접적인 방법을 동원하였다. 본 논문에서는 기존의 방법들을 정리해보고 여기에 조정변수를 사용하는 새로운 방법을 추가하여 제시하였으며 이러한 방법들의 효율성을 실제의 예를 통해 보여 주고 있다.

1. 다차원 자료의 구조탐색을 위한 고전적 방법

대부분의 현실 자료는 다차원이며 이들은 테이블 형식(행렬 형식)으로 기록된다. 이 테이블로부터 각 관측값들의 성격을 상호 비교하거나 각 변수들의 상호 관계 등을 조사하는 과정은 매우 어렵거나 불가능하다. 통계학자들은 이러한 문제를 해결하기 위해 여러가지 통계적인 이론들을 제시할 뿐 아니라 이러한 이론들의 효율을 조사하기 위해 많은 노력을 계속하고 있다. 여기서는 다차원 자료를 그림으로 표현하므로서 분석자가 직관적인 판단에 의해 자료의 구조를 탐색하는 방법을 알아보고자 한다.

자료의 구조를 파악하기 위한 가장 쉬운 방법은 일변수 통계량을 그림으로 표현하기 위해 개발된 도형을 각 변수에 적용시키는 것이다. 보편적으로 이용되는 방법들을 열거하면 막대그래프, 줄기-잎 그래프, 누적분포함수, 박스 플롯, 라인 플롯(일차원 직선에 자료를 투사시키는 방법), q-q 플롯 등이 있다. Easton과 McCulloch(1990)는 다차원 자료가 가정한 분포를 따르는가를 알아보기 위해 q-q 플롯을 다차원으로 확대시키는 연구를 하였다. 또한 허문열(1995a, b)에서 제시되고 있는 FEDF도 일변수 자료의 구조를 파악하는 데 유용하다. 이들 방법은 적용하기가 쉽고 누구든지 곧 이해할 수 있는 장점이 있지만 변수들 간의 상호 관계를 알 수 없는 것이 기본적인 제한 점이다.

두 변수 사이의 상호 관계를 탐색하는 데 많이 이용하고 있는 산점도를 다차원 자료의 구조를 탐색하기 위해 확장시키는 방법에 대해서 많은 연구가 이루어졌다. 이 중에 대표적인 것이 Carr 등(1986, 1987)에 의해 제시된 것으로서 산점도의 각 점을 glyph로 나타내는 것이다. 즉, 산점도의 각 점에 방향을 갖는 선을 첨가하면 4차원 자료를 표시할 수 있다. 이 아이디어는 자료를 비슷한 성격으로 분류하는 데 유용하게 이용될 수 있다. 그러나 자료의 양이 많아지면 그림이 매우 복잡해지므로 유용성이 급격히 떨어지게 된다. 또 겹치는 점이 많은 경우에도 문제가 된다. 2차원 산점도의 경우 겹치는 점을 해결하는 방법으로 Cleveland 와 McGill(1984)은 sunflower 플롯을 제안

1) 본 논문은 한국과학재단의 94년도 핵심전문연구과제 지원에 의해 이루어졌음
2) (110-745) 서울특벽시 종로구 명륜동 3가 53번지, 성균관대학교 통계학과

하였으며 Schilling과 Watkins(1994)에 의해 보완되었다.

산점도를 p-차원 자료에 확장하는 또 다른 방법으로는 p(p-1)/2 쌍의 산점도를 행렬로 표현하는 것이다. 이 아이디어는 Stuetzle (1988)이 그들이 개발한 Plot-Windows 패키지에서 구현하였으며 S-plus에서 다차원 자료를 분석하는 기본이 되고 있다. 그러나 산점도 행렬만 보아서는 다차원 자료에 대한 정확한 정보를 파악할 수 없다. 이것은 마치 p개의 인자를 갖는 인자분석 모델에서 2개의 교호작용 까지만 고려하는 것과 마찬가지이다.

이상의 방법들은 일차원이나 이차원 자료를 파악하기 위해 개발된 도구를 다차원 자료의 탐색을 위해 확장시킨 것이다. 그러나 다차원 자료를 파악하는 목적만으로 개발된 방법들도 많이 있다. 이들 중에서 대표적인 것들은 Chernoff의 얼굴(1973)로서 사람 얼굴의 눈, 코, 입, 머리카락, 귀, 턱 등을 이용하여 다차원 자료의 구조를 파악하는 것이다. 또 Wakimoto(1978)는 반원 위에 다차원 자료를 갖는 각 관측값이 하나의 점으로 표시되는 별자리 좌표를 고안하였다. 별자리 좌표에서는 같은 성격(각 변수의 입장에서)을 갖는 관측값들은 반원상에서 뭉쳐서 나타난다. 따라서 이 그림을 이용하면 집락분석과 같은 다차원 자료분석이 가능하다(이경미,1993). Statgraphics 팩키지 등에서 사용하는 star 플롯은 각 관측값을 하나의 별로 표현하고 있다. 이 경우 변수의 수는 별이 빛날 때 나타나는 꼭지의 수에 해당된다. 이를 이용하면 변수의 성격들이 유사한 관측값들은 유사한 모양의 별로 표시된다. 따라서 이를 이용하면 직관적으로 자료들을 그룹으로 분류할 수 있다. 이 외에도 Andrews(1972)는 후리에 시리즈 표현 방법을 이용하여 다차원 자료를 표현하였으며 Inselberg(1985)는 평행좌표 플롯이라는 개념을 도입하여 다차원 자료를 표현하였다. 그러나 이들 마지막 두 방법은 관측값의 수가 많아질 때 거의 효용성을 잃어버린다.

형상의 차원이 3개 이상이 되면 현실적으로 이 형상을 2차원 평면에 그대로 표현하는 것이 불가능해진다. 마찬가지로 대부분 현실적인 문제에서 자료의 변수는 3개 이상이므로 이들 자료의 내용을 2차원 평면에 그대로 표현하는 것이 불가능하다. 따라서 통계학자들은 앞에서 설명한 바와 같은 간접적인 방법을 통하여 고차원 자료의 구조를 파악하려고 노력하였다.

컴퓨터 그래픽스의 기술이 발달하므로서 다차원 자료의 구조를 2차원 상에서 어느 정도 파악할 수 있는 방법이 개발되기 시작하였다. 이 중에 대표적인 것이 Becker 등(1988)에 의해 제안된 동적그래픽스 방법으로서 이는 주로 빗질(brushing)과 연결(linking)에 의해 다차원 자료의 구조를 파악하는 것이다. 2절에서는 이러한 방법들에 대해 설명하고자 한다. 또한 자료의 회전을 통해 보이지 않는 부분을 탐색하는 방법이 있으며 이것의 대표적인 방법이 Asimov(1985)가 제시한 grand-tour이다. 3절에서는 자료의 회전을 통해 숨겨져 있는 다차원 구조를 파악하는 방법에 대해 설명한다. 이 외에 Tukey (1973) 등은 조정 변수를 사용하므로서 (n+1)차원 자료의 구조파악을 시도하였다. 4절에서는 Tukey의 조정변수 개념을 보완한 누적조정변수의 도입을 설명하고 실제 자료를 사용하여 이 방법의 효율성을 보이고자 한다.

2. 빗질과 연결에 의한 다차원 자료의 구조탐색

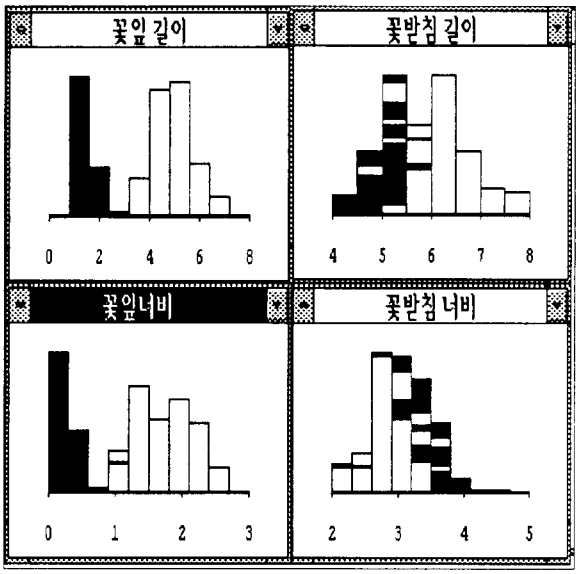
일차원 자료나 2차원 자료를 표현하기 위해 만든 여러가지 방법들은 이들을 어떠한 방법으로 나열하더라도 3차원 이상의 자료를 완벽하게 표현할 수 없다. 그러나 동적 그래픽스 방법을 적용한다면 이를 어느 정도 보완할 수 있다.

동적 그래픽스의 중요성과 구현 방법, 그리고 이를 위한 계산 환경 등 전반적인 내용에 대해서는 Becker 등(1988)이 편집한 책자에서 자세히 설명하고 있다. 이들은 동적그래픽스의 특성을 '직접조작과 즉각적인 실현'이라고 하였다. 여기서 '직접조작'이라는 것은 화면상에서 마우스 등을 이용하여 도형의 특성을 직접 조작하는 것을 말하며, 이렇게 직접 조작한 내용의 결과는 화면에 즉각적으로 나타나야 한다. 예를 들어 2차원 자료를 이용하여 산점도를 그리고 여기에 최소제곱 적합직선을 그려넣었다고 하자. 이 때 관측값 하나를 마우스로 움직이면 이와 동시에 적합직선이 다시 만들어지고, 또 이 직선이 다시 그려지도록 하였다면 이는 동적그래픽스 방법에 의해 구현되었다고 할 수 있다. XLISP-STAT [17]에 이를 구현해 놓은 예제가 있으나 여기서는 자료를 움직이는 시점과 동시에 적합직선이 그려지는 것이 아니고 자료를 움직이고 난 후에 적합직선이 다시 그려진다.

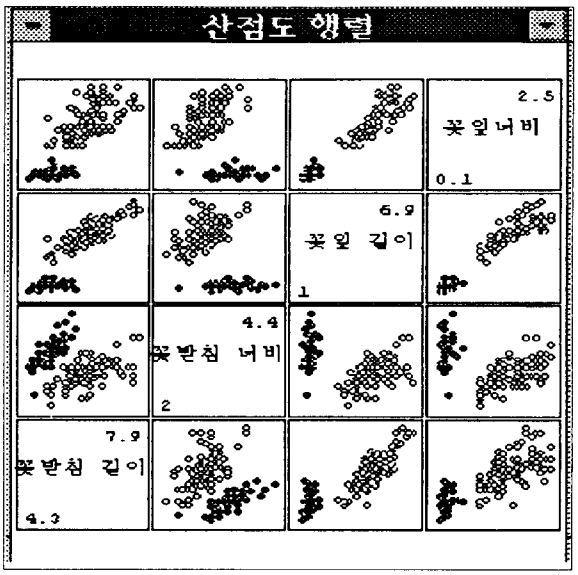
동적그래픽스의 구현 방법은 도형에 나타난 관측값을 빗질(brsuhing)에 의해 선택하고 이들을 다른 도형에 나타난 관측값들과 연결(linking)하는 것이 기본이다. 또한 마우스를 사용하여 자료를 변형시키거나 삭제할 수 있어야 한다. 예를 들어 2차원 자료를 분석하기 위해 각 변수별로 막대그래프를 그렸다고 하자. 이 두 개의 막대그래프만 가지고는 두 변수간의 관계를 알 수 있는 방법이 없다. 그러나 첫 번째 막대그래프에서 중간 이상의 관측값들을 택하고 이렇게 선택된 관측값이 두 번째 막대그래프 어디에 나타나는가를 연결하여 찾아본다면 두 변수들의 연관 관계를 간접적으로 파악할 수 있다. 동적그래픽스의 구현은 일차원 도형들만으로도 가능하지만 일차원 도형과 이차원 도형들을 혼합하여 사용할 수도 있다.

다차원 자료분석에서 자주 이용되는 분꽃 자료를 사용하여 4개의 막대그래프와 산점도 행렬을 그린 것이 <그림 1>과 <그림 2>에 나타나 있다. 이 그림들에서 검정색으로 칠해진 꽃들은 꽃잎 길이가 2 이하인 것들이다(꽃잎 길이는 최소가 1이고 최대가 6.9이다). 먼저 4개의 막대그래프를 살펴보면 꽃잎 길이가 작은 꽃들은 모두 꽃잎 너비가 작은 것으로 나타났다. 꽃받침의 경우, 꽃잎 길이가 작은 꽃들의 꽃받침 길이는 비교적 작지만 반대로 꽃받침 너비는 큰 꽃들이 많은 것을 알 수 있다. 산점도 행렬을 살펴보면 재미있는 현상을 발견할 수 있다. 즉, 전체 150개의 꽃들이 꽃잎 길이 또는 꽃잎 너비를 기준으로 하여 완전히 두 개의 그룹으로 분리되어 있는 것을 알 수 있다. 이러한 정보를 4개의 일차원 막대그래프에서 획득하는 것은 매우 어렵다.

산점도 행렬에서 대각 요소에 막대그래프를 그려놓고 꽃잎 막대그래프에서 길이가 작은 꽃들만 택하여 이를 다른 그림과 연결하므로서 다차원 자료의 구조를 파악하려는 시도는 이미 Stuetzle 등(1986)에 의해서 연구된 바 있으며, 이 과정은 이미 S-plus에 구현되어 있다. 또 Unwin (1993)은 막대그래프 대신 박스 플롯을 이용하였다. 그러나 <그림 1>에서 볼 수 있는 바와 같이 막대그래프의 각 기둥에 속해 있는 여러 개의 관측값들은 서로 구분이 되지 않아 해석에 어려운 점이 많다. 이를 보완하는 방법으로 Stueztle(1986)은 막대그래프의 각 기둥을 점(dot)로 표시하였으나 이렇게 한다고 하여 근본적인 문제가 해결되지는 않는다. 이에 허문열(1995b)은 FEDF라는 개념을 도입하였다. FEDF(Flipped Empirical Distribution Function)는 EDF를 중앙값까지 그리고, 중앙값을 넘는 관측값에 대해서는 y=0.5를 중심으로 하여 EDF를 꺾어놓았다. 이를 이용하면 각 관측값이 막대그래프 처럼 중복되어 나타나는 일이 없을 뿐만 아니라 이 그래프로부터 해당 변수의 분포형태가 대칭인가를 즉시에 판단할 수 있고 간단한 통계량을 직접 얻을 수가 있다. 더욱이 각 관측값의 위치를 직접 찾아갈 수 있기 때문에 동적그래픽스에 매우 유용하게 이용된다. FEDF를 이용하면 이 외에도 여러가지 유용한 정보를 얻을 수가 있다.



<그림 1>
분꽃 자료의 4변수를 각각 막대그래프로 나타내고 꽃잎 길이가 작은 꽃들만 택하여 이를 표시한 결과



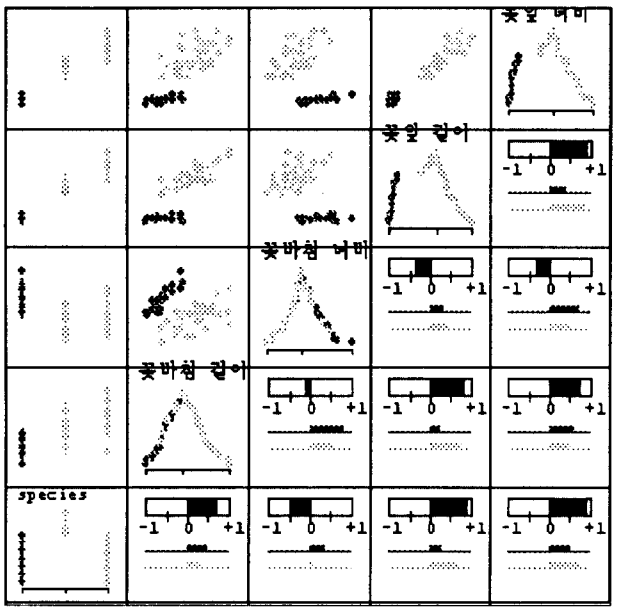
<그림 2>
분꽃 자료의 4변수를 산점도 행렬로 나타내고 꽃잎 길이가 작은 꽃들만 택하여 표시한 결과

<그림 2>에 나타나 있는 산점도행렬은 대각 요소에 해당 변수의 이름과 최소값, 최대값이 나타나 있고 우측 하단에는 좌측 상단의 산점도와 대칭인 형태가 주어졌다. 산점도 행렬의 대각요소에 각 변수의 FEDF를 넣고 우측 하단 행렬에는 산점도 행렬 대신 해당 두 변수들에 대한 연관성 척도를 제공한다면 더욱 효용성이 높아질 것이다. 이를 고려하여 Huh(1995b)는 <그림 3>과 같은 FEDF-산점도행렬을 제안하였다. 우측하단의 각 셀에는 몇 개의 막대가 그려져 있다. 첫 번째 막대는 해당 두 변수간의 상관계수이고 두 번째 부터 나타나 있는 막대는 조건부 상관계수이다. 조건부 상관계수는 연결되어 있는 관측값들끼리의 상관계수를 의미한다. 여기서 상관계수의 계산은 순위 상관계수를 적용하였다. FEDFD와 FEDF-산점도행렬에 대한 자세한 내용은 Huh (1995b)에 나와 있다.

3. 자료의 회전을 통해 다차원 자료의 구조를 파악하는 방법

여기서는 두 가지 측면에서 다차원 자료를 탐색하는 방법을 생각해 보기로 한다. 첫 번째 방법은 자료를 회전시킴으로서 구조를 파악하는 방법이고, 두 번째 방법은 n차원 도형에 새로운 조정변수를 추가시키므로서 (n+1)차원 자료를 탐색하는 방법이다.

먼저 자료를 회전하므로서 자료의 구조를 탐색하는 방법을 살펴보기로 한다. 다차원 자료를 2차원에 표현하면 어떠한 방법을 사용하더라도 많은 정보를 상실하게 된다. 예를 들어 서울의 남산을 2차원에 지면(또는 화면)에 표현하는 것에 대해 생각해 보자. 보편적인 방법은 입면도와 평면도,



<그림 3>
분꽃 자료의 4 변수를 산점도 행렬로 나타내고 대각 요소에는 각 변수의 FEDF를 그려넣었다. 또 우측하단의 각 셀에는 상관계수를 나타내는 막대를 그렸다.

그리고 측면도를 사용한다. 그러나 이 3개의 2차원 도형만 가지고는 남산의 정확한 구조를 파악할 수 없다. 이는 3차원 사람들은 2차원 도형까지만을 한 눈에 알아볼 수 있기 때문이다. 일반적으로 (n+1)차원 사람은 n차원 도형 까지만 한 눈에 알아볼 수 있다.

이제 3차원 도형인 남산을 여러 각도에서 돌려가며 연속적인 2차원 도형을 만든다면 3차원 남산의 형상을 2차원에서 가상적으로 볼 수 있을 것이다. 이러한 기법을 통계적인 자료분석에 적용시켜본 것이 S-plus와 XLISP-STAT등의 패키지에서 볼 수 있는 spin-plot이다. 도형을 회전시킬 때 회전 축은 자료가 몇 차원이냐에 따라 달라진다. 2차원 도형인 경우 회전축은 점(원점)이다. 3차원인 경우 회전축은 선이다. 그러나 4차원인 경우 회전축은 평면이다. 또 5차원인 경우 회전축은 3차원 이라고 유추할 수 있다. 여기에 대한 논리적인 배경은 1차원 직선은 0차원 점을 움직이므로서 만들어지고, 2차원 평면은 직선을 움직이므로서 만들어지며, 3차원 공간은 2차원 평면을 움직이므로서 만들어진다고 생각할 수 있다. 이렇게 유추하면 4차원 초평면은 3차원 공간을 움직이므로서 만들어지고, 5차원 초평면은 4차원 초평면을 움직이므로서 만들어진다. 따라서 2차원 평면을 회전시킨다는 것은 이 평면을 만드는 직선을 직선 상의 한 점을 중심으로 하여 회전하고 이를 움직이므로서 이루어진다고 생각할 수 있다. 또 3차원 공간을 회전하려면 이 공간을 만드는 2차원 평면을 평면상의 한 직선을 축으로 하여 회전하고 이를 움직여서 만드는 것으로 생각할 수 있다. 이렇게 유추할 때 4차원 초평면의 회전축은 이 초평면을 만드는 4개의 축 중에서 2개의 축으로 이루어진 2차원 평면이 된다.

3차원 도형을 2차원 평면에 표현할 때 일반적인 도형인 경우 원근법과 은선처리, 그림자 처리 등을 통해 3차원적인 공간을 상상할 수 있다. 그러나 3차원 자료를 2차원 평면상의 점으로 표현하는 경우 세밀한 관찰력과 훈련 없이는 이를 이해하는 것이 어렵다. 예를 들어 훈련되지 않은 사람인 경우 산점도를 통해 2차원 자료를 파악하는 것 자체가 어려운 것을 생각하면 이 문제를 이해

할 수 있다.

3차원 자료의 구조를 2차원에서 파악할 수 있도록 하기 위해 많이 사용하는 방법은 두 가지이다. 첫 번째 방법은 시점에서 멀리 떨어져 있는 점은 작게(또는 희미하게) 표시하고 가까울 수록 크게(또는 밝게) 표시하는 것이다. 두 번째 방법은 이미 앞에서 설명한 바와 같이 도형을 회전시키는 것이다. XLISP-STAT의 spin-plot은 이 두 가지를 동시에 한 것이다.

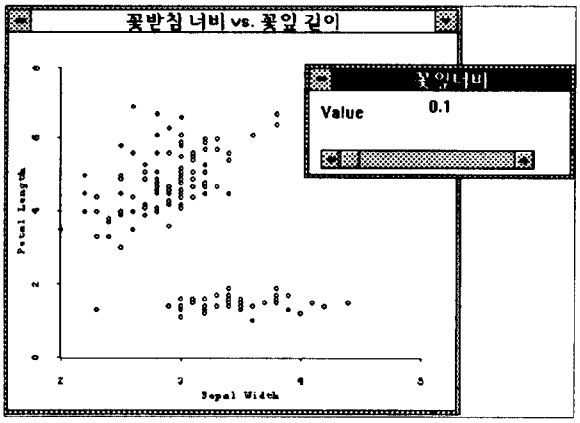
4차원 이상의 자료인 경우 Asimov는 grand-tour라는 방법을 도입하여 다차원 자료를 탐색하려고 시도하였다. 이것은 문자 그대로 다차원 자료를 여기 저기 구석 구석 찾아가며 무엇이 있는가를 살펴본다는 아이디어를 수학적으로 구현한 것으로서 XLISP-STAT의 tour-plot 함수가 이 기능을 제공하고 있다. 그러나 다차원 자료의 구조를 grand-tour로 파악한다는 것은 매우 난해하다. 이 외에도 다차원 자료를 2차원 화면에 표현하는 여러가지 방법이 제시되었지만 자료가 4차원 이상이 되면 이를 2차원 화면에 직접 표현하여 그 구조를 파악한다는 것이 사실상 불가능하다고 할 수 있다.

4. 누적조정변수를 사용한 다차원 자료의 구조탐색

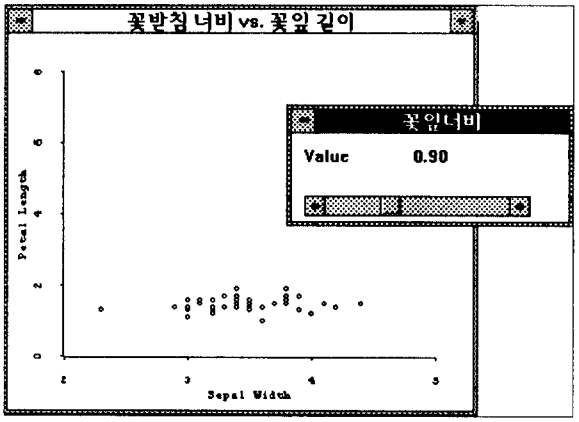
이제 방향을 바꾸어서 다시 2차원 산점도를 생각해 보자. 산점도의 변수와 별도로 또 하나의 변수를 두어 이 변수의 값이 변함에 따라 여기에 대응하는 관측값들을 산점도에 연결시키면 3차원 자료를 파악할 수 있다. 이 방법은 앞 절에서 설명한 동적 그래픽스의 기본적인 아이디어를 이용한 것으로서 이미 Tukey와 그의 동료들에 의해 1973년에 PRIM-9 이라는 프로그램을 통해 컴퓨터로 구현된 바 있다. 그러나 PRIM-9의 문제점은 관측값의 수가 작으면 산점도를 조정하는 변수가 어떤 값을 가질 때 여기에 대응해서 나타나는 관측값의 수가 없거나 적다. 따라서 현실적인 의미를 상실한다. 예를 들어 조정하는 변수의 범위가 100에서 200 사이의 값을 갖는다고 하고 관측값의 수가 20개라고 하자. 변수의 값을 100에서 200으로 1씩 증가시키면서 변화시킨다고 하면 이 값에서 관측값의 수가 하나도 없는 경우가 대부분이다. 이를 보완하는 방법으로는 조정변수가 갖는 값의 범위를 현재 관측값이 갖는 최소값에서 최대값까지로 하여 조정 변수가 현재 갖는 값 까지 누적된 관측값들을 산점도에 연결하는 것이다. 더욱이 처음에 시작하는 단계에서는 산점도를 비워놓고 조정변수의 값이 증가함에 따라 여기에 속하는 관측값들을 산점도에 추가하여 그리면 자료의 구조 파악이 더욱 쉬워진다.

붓꽃 자료를 사용하여 이상의 내용을 설명한다. <그림 4>에 꽃받침 너비(sepal width)와 꽃잎 길이(petal length)의 산점도가 나타나 있다. 또 dialog 창에는 꽃잎 너비(petal width)의 값을 마우스로 조정할 수 있는 막대가 주어져 있다. 이 막대를 마우스를 이용하여 우측으로 옮겨가면 조정변수인 꽃잎 너비의 값이 증가하고 여기에 속하는 관측값들만 산점도에 그려진다. <그림 5>에 꽃잎 너비가 최소값(0.1) 부터 현재 마우스가 지적하고 있는 값(0.9) 까지의 관측값들이 산점도에 나타나 있다. 다시 마우스를 우측으로 2.0 까지 이동하고 여기까지 속한 관측값들을 산점도에 연결시킨 것이 <그림 6>에 나타나 있다. 이 세 그림을 통해서 우리는 꽃잎 너비가 커짐에 따라 붓꽃 자료가 두 개의 그룹으로 구분되는 것을 명확하게 파악할 수 있다.

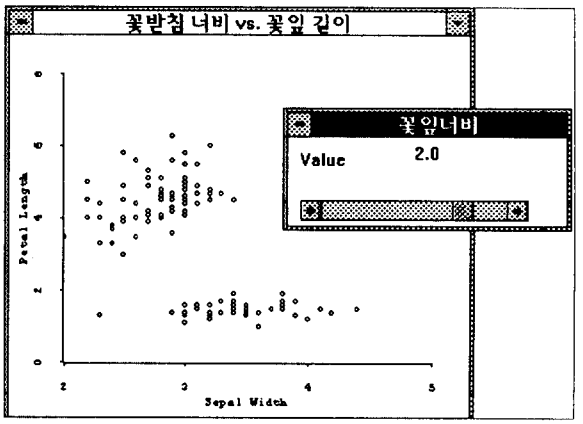
산점도 행렬 대신 다차원 직교좌표나 grand-tour를 이용할 수도 있다. 다차원 직교좌표를 이용하는 경우 3차원 자료까지는 무난히 이해할 수 있으나 차원의 수가 이를 넘어서면 그림이 난해해



<그림 4>
분꽃 자료에서 꽃잎 길이와 꽃받침 너비의 산점도를 그리고 이와 별도로 꽃잎너비를 dialog 창에 만들어 이를 이용하여 산점도에 나타나는 관측값을 조정한다

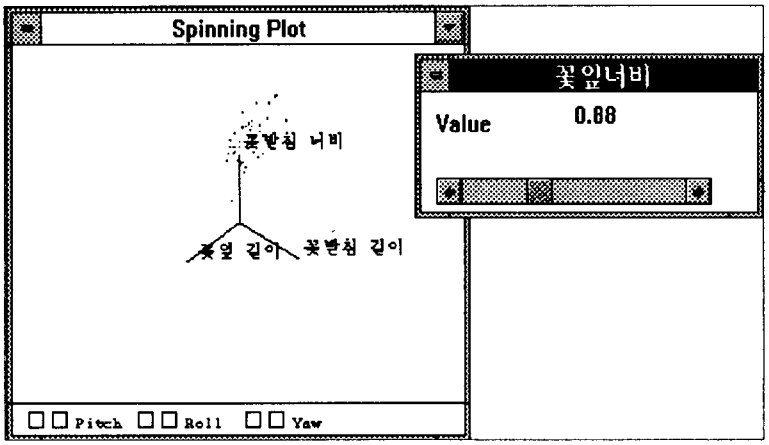


<그림 5>
분꽃 자료 중에 dialog의 조정변수로 꽃잎 너비를 최소값 0.1부터 시작하여 0.9까지 증가시키면서 여기에 속한 관측값들이 산점도에 어떻게 나타나는가를 알아본 결과 꽃잎 길이는 작고 꽃받침 너비는 좌우로 퍼져서 나타나는 것을 알 수 있다

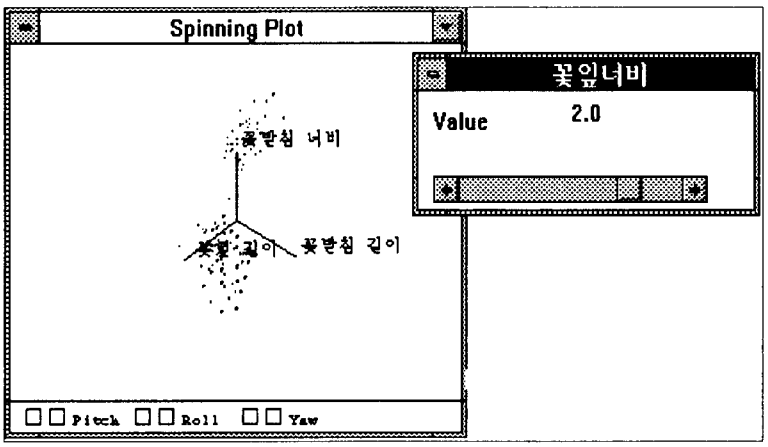


<그림 6>
분꽃 자료 중에 dialog의 조정변수로 꽃잎 너비를 2.0까지 증가시켜 보면 꽃잎 길이는 점점 증가하는 반면 꽃받침 너비는 줄어드는 것을 알 수 있다

진다. 분꽃 자료를 사용하여 예를 들어 본 것이 <그림 7>과 <그림 8>에 나타나 있다. 조정변수로 꽃잎 너비를 사용하여 dialog창에 나타내었고, 나머지 3 변수를 3차원 직교좌표상에 표현하였다. 이 그림에 의하면 꽃잎 너비가 작은 경우 꽃잎 길이와 꽃받침 길이는 작은 부분에서 랜덤하게 나타나고 꽃받침 너비는 긴 부분에서 랜덤하게 나타나는 것을 알 수 있다. 그러나 꽃잎 너비가 1.0을 넘어서면서 꽃잎 길이와 꽃받침 길이가 긴 부분에서, 그리고 꽃받침 너비는 작은 부분에서 랜덤하게 나타나는 것을 알 수 있다. 다시 말하여 분꽃 자료는 꽃잎 너비에 의해 두 그룹으로 구분할 수 있는 것을 이 그림을 통해 즉시 알 수 있다.



<그림 7>
분꽃 자료 중에 dialog의 조정 변수 꽃잎 너비를 0.88 까지 증가시켜보면 꽃잎 길이와 꽃받침 길이는 작은 부분에서 랜덤하게 움직이지만 꽃받침 너비는 큰 곳에서 움직이는 것을 알 수 있다



<그림 8>
분꽃 자료 중에 dialog의 조정 변수 꽃잎 너비를 계속 2.0까지 증가시켜보면 꽃잎 너비가 1.0 근처에서 꽃잎 길이와 꽃받침 길이가 갑자기 커지면서 꽃받침 너비는 작아지는 것을 알 수 있다. 따라서 분꽃 자료는 꽃잎 너비를 1.0을 기준으로 하여 두 개의 그룹으로 분류할 수 있는 것을 알 수 있다

부 록

본 논문에 나타나는 그래프는 모두 XLISP-STAT 을 사용하여 작성하였으며 이를 만들어주는 프로그램은 다음과 같다. 다만 그림 1, 2, 3은 간단하여 프로그램 코드를 제시할 필요가 없어 생략한다.

```
;;;
;;; load iris data: 4 by 150
;;;
(def data (load "iris"))
;;;
;;; draw scatterplot with control-variable
;;;
(defun scatter-with-control (s-var c-var)
"s-var: list of variable numbers for scatterplot
 c-var: list of variable numbers for control variable"
  (let* ((x (first s-var))
         (y (second s-var))
         (x-data (elt iris x))
         (y-data (elt iris y))
         (s-data (elt iris c-var))
         (x-label (elt varnames-e x))
         (y-label (elt varnames-e y))
         (s-label (elt varnames-h c-var))
         (title (format nil "~s vs. ~s"
                        (elt varnames-h x) (elt varnames-h y)))
         (xy (send graph-proto :new 2 :title title)))
    (send xy :range '(0 1)
          (list (min x-data) (max x-data))
          (list (min y-data) (max y-data)) :draw nil)
    (send xy :x-axis t t 4 :draw nil)
    (send xy :y-axis t t 4 :draw nil)
    (send xy :variable-label '(0 1) (list x-label y-label))
    (flet
      ((xy-plot (val)
                (let* ((sel (which (>= val s-data)))
                       (x-sel (select x-data sel))
                       (y-sel (select y-data sel)))
                  (send xy :clear :draw nil)
                  (send xy :add-points (list x-sel y-sel) :draw nil)
                  (send xy :adjust-to-data))))
      (interval-slider-dialog
       (list (- (min s-data) 0.1) (+ (max s-data) 0.1))
       :action #'xy-plot
```

```
        :title s-label))))

;;;
;;; draw spinning plot with control-variable
;;;
(defun spinning-with-control (s-var c-var)
"s-var: list of variable numbers for spinningplot
 c-var: list of variable numbers for control variable"
  (let* ((s-data (elt iris c-var))
         (s-label (elt varnames-h c-var))
         (xy (spin-plot (select iris s-var)
                        :variable-labels (select varnames-h s-var))))
    (flet
      ((xy-plot (val)
                (let* ((sel (which (>= val s-data)))
                       (data (mapcar #'(lambda (x)
                                         (select x sel)) (select iris s-var))))
                  (send xy :clear :draw nil)
                  (if (first data) (send xy :add-points data :draw nil))
                  (send xy :adjust-to-data))))
       (interval-slider-dialog
        (list (- (min s-data) 0.1) (+ (max s-data) 0.1))
        :action #'xy-plot
        :title s-label))))

(scatter-with-control '(1 2) 3)
(spinning-with-control '(0 1 2) 3)
```

참고문헌

[1] 이 경 미 (1993). 동적 그래픽스에 의한 군집분석, 석사학위 논문, 성균관대학교.

[2] 허 문 열 (1995a). 컴퓨터 그래픽스에 의한 이원분산분석, 「응용통계연구」, 제 8권 1호, 75-87.

[3] Andrews, D.f (1972). Plots of High Dimensional Data, *Biometrics,* Vol. 28, 125-136.

[4] Asimov, D (1985). The Grand Tour: a Tool fo Viewing Multidimensional Data, *SIAM Journal of Scientific and Statisticcal Computing,* Vol. 6, 128-143.

[5] Becker, R. A., Cleveland, W. S., and Wilks, A. R. (1988). Dynamic Graphics for Data Analysis, *Dynamic Graphics for Statistics*, edited by Cleveland, W., and McGill, M. Wadsworth & Brooks.

[6] Carr, D.B.,Nicholson, W.L., Littlefield, R.J., and Hall, D.L. (1986). Interactive Color Display

Methods for Multivariate Data, *Statistical Image Processing and Graphics*, Wegman, E.J. and Depriest, D.J. (eds.), 215-250, Marcel Dekker, New York.

[7] Carr, D.B., Nicholson, W.L., Littlefield, R.J., and Hall, D.L. (1987). Scatterplot Matrix Techniques for Large N, *The Journal of the American Statistical Association*, Vol. 82, No. 398, 424-436.

[8] Cleveland, W.S., and McGill, R. (1984). The Many Faces of a Scatterplot, *The Journal of the American Statistical Association*, Vol. 79, 807-822.

[9] Easton, G. S., and McCulloch, R. E. (1990). A Multivariate Generalization of Quantile-Quantile Plots, *The Journal of the American Statistical Association*, 376-386.

[10] Fisher, R.A. (1936). The use of multiple measurements in Taxonomic problems, *The Annals of Eugenics*, Vol. 7, 179-184.

[11] Huh, Moon Yul (1995b). Dynamic Graphics with FEDF, to appear in *The Journal of Computational and Graphical Statistics*, Vol. 4, No. 4, 1-9.

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

[13] Schilling, M.F., and Watkins, A.E. (1994). A Suggestion for Sunflower Plots, *The American Statistician*, Vol. 48, No. 4, 303-305.

[14] "PRIM-9" (1973). Produced by Stanford Linear Accelerator Center, Stanford, Ca. (John Tukey, lecturer) Bin 88 Productions. (Film).

[15] Scott, David W. (1992). *Multivariate Density Estimation*, Addison-Wesley.

[16] Stuetzle, Werner (1988). Plot Windows, *Dynamic Graphics for Statistics*, edited by Cleveland, W., and McGill, M. Wadsworth & Brooks.

[17] Tierney, Luke (1990). *LISP-STAT*, John Wiley & Sons.

[18] Unwin, Anthony (1993). In the kingdom of the blind no body uses interactive graphics, *Statistische Programmiersprachen und Interaktive Datenanalyse, Internationale Biometrische Gesellschaft*, Oct. 8-11, 1993, Stift Keppel, Germany.

[19] Wakimoto, K., and Taguri, M. (1978). Constellation Graphical Methods for Representation of Multidimensional Data, *The Annals of the Institute of Statistical Mathematics*, A30, 97-104.