머리말

1995년 고려대학교 통계연구소 강의 총서 19권째로 출간된 『SAS 다차원척도법』은 다차원척도법을 정리해 보고자 하였던 동기 부여 차원에서 이루어졌으나 보면 볼수록 부족함이 많았다. 사실 이 책의 전면 개정과 보완이 필요하다고 생각은 하였지만, 다차원척도법에 대한 지식이 부족하여 많이 망설였다. 그리고 이를 극복하기 위하여 대학원에서의 다차원척도법 강의와 학생들과의 세미나를 통하여 조금씩 관련 자료를 준비해 왔다.

2012년에 연구년으로 아이오와 주립대학에 오면서도 다차원척도법의 완성도를 높이고 싶은 마음은 있었지만 몇 권의 졸작들을 개정하는 작업의 범주에 넣을 엄두를 내지는 못하였다. 그런데 의외로 목표하였던 나머지 개정 작업이 순조롭게 진행되면서 거닐고 싶은 길이 있는 공원을 산책하듯 편안한 마음으로 본 책인『다차원척도법의 산책』을 집필할 수 있게 되었다.

이 책은 『SAS 다차원척도법』 의 큰 흐름을 따르면서 내용과 사례를 보완하였고 구성을 조금 바꾸면서 다양한 자료를 추가하였다. 2장의 비유사성과 유사성, 6장의 베이즈 다차원척도법, 그리고 7장의 비선형 다차원척도법이 여기에 해당한다. 특히, 응용의 묘미를 위해 주된 프로그램은 SAS/MDS, SAS/IML과 R이 사용되었고, SAS/STAT 9.2판의 ODS Graphics 문은 6.07판의 PROC PLOT 문의 번거로움을 덜어주었다. 현재 R에서 계량형과 비계량형 그리고 비선형 다차원척도법을 위한 cmdscale 함수, isoMDS 함수와 sammon 함수가 제공되고 있다.

다차원척도법(MDS: multidimensional scaling)이란 다차원 공간에서 개체 간의 거리를 나타내는 자료로부터 그들의 유사성 또는 비유사성을 저차원 공간에 기하적으로 나타내어 그들의 관계를 탐색적으로 살펴보는 다변량 그래프적 기법이다. 1950년 이후 60년 이상의 역사를 갖고 있는 이 기법은 정치학, 경제학, 심리학, 사회학, 지리학, 교육학 등 여러 분야에 걸쳐 다양하게 응용되고 있다.

이 책의 구성과 내용을 요약하면 다음과 같다.

- 1장: 다차원척도법의 정의와 사례를 제시하고 역사적 배경과 이론적 배경을 살펴본다. SAS/MDS를 활용한 사례분석은 다차원척도법의 개념을 잡는 데 도움이 될 것이다.
- 2장: 개체 간과 변수 간의 비유사성과 유사성을 정의하고 이를 측정할 수 있는 측도에 대하여 소개한다. 비유사성과 유사성 상호 간의 변환을 정의하고 비대칭인 비유사성과 유사성에 대한 변환도 소개한다.
- 3장: 이원 다차원척도법인 계량형과 비계량형 다차원척도법의 토거선과 크루스칼-세퍼드 알고리즘을 각각 소개한다. 특히, SAS/MDS와 R-프로그램(cmdscale 함수와 isoMDS 함수)에서 제공하는 두 다차원척도법을 비교하고 있다.
- 4장: 각 주체에 대해 고려한 비유사성행렬에 적용하는 삼원 다차원척도법의 이론과 이를 위한 INDSCAL 알고리즘을 소개한다.
- 5장: 다차원척도법에서 분포를 가정하고 최대우도추정법을 적용한 최대우도 다차원척도법의 이론과 SAS/MDS에서 이와 유사한 결과를 제공하는 로그램과 사례를 소개한다.
- 6장: 계량형 다차원척도법에서 베이즈 모형에 바탕을 두고 MCMC 알고리즘을 적용한 베이즈 계량형 다차원척도법을 소개한다. 이를 위한 bmds 함수를 활용한 R-프로그램을 사용하고 있다.
- 7장: 계량형 다차원척도법에서 비선형사상 알고리즘을 위한 새먼사상을 적용한 비선형 다차원척도법을 소개하고 sammom 함수를 활용한 R-프로그램도 제시하고 있다.

- 부록 1: SAS MDS 절차를 소개하며 특히, PROC MDS 문법을 정리하고 요약한다. 범용 PROC ALSCAL/MLSCALE과의 간단한 비교도 하였다.
- 부록 2: 자료 1-자료 13의 13가지 자료를 소개하고 있다. 이는 본문과 [연습문제]에서 비교적 큰 자료 위주로 파일과 그 이름을 정리하여 제공하고 있다.
- 부록 3: 각 장별과 부록 1에서 사용된 프로그램의 파일을 정리 요약하고 있다.

3~7장에서 제시되는 다차원척도법의 알고리즘에 대한 이론적 설명은 어려우면 넘어가도 좋다. 그러나 예제를 통해 제시하는 프로그램과 실행결과 그리고 다차원척도 그림에 대한 설명은 자세히 읽어보기를 바란다. 참고로 저자의 홈페이지(home.pusan.ac.kr/~yschoi)에서 자료와 프로그램을 내려 받을 수 있다.

6장의 베이즈 다차원척도법을 위한 참고 프로그램과 자료를 알려주신 이화여자대학교 오만숙 교수께도 감사의 인사를 드린다. 그리고 연구년 동안 좋은 환경을 제공하신 아이오와 주립대학 PPSI의 소장 Richard Spoth 박사와 부소장 신충렬 박사, Cleve Redmond 박사께 감사드린다. 끝으로 헌신적인 초고 읽기를 통하여 오류를 지적한 박사 과정의 이보희와 지도 학생들 그리고 마지막까지 bmds 함수를 활용한 R-프로그램의 오류를 찾아낸 이수기 군에게도 고마움을 전한다.

2013년 아이오와 에임스의 눈을 즐기며
아이오와 주립대학 PPSI 연구실 2442호
최 용 석

차 례

1장 다차원척도법의 이해

1.1 다차원척도법이란 · 2

1.1.1 [사례 1] 우리나라 도시 간 철도거리 · 4

1.1.2 [사례 2] 우리나라 14개 경제관련기관의 경제 전망 · 6

1.1.3 [사례 3] 12개 국가의 유사성에 대한 평가 · 11

1.1.4 [사례 4] 15군데 신체 부위의 비유사성 · 13

1.1.5 [사례 5] 15가지 레크리에이션의 비유사성 · 15

1.2 다차원척도법의 역사적 배경 · 18

1.3 다차원척도법의 이론적 배경 · 20

1.3.1 측정수준 · 20

1.3.2 스트레스 · 22

1.3.3 차원의 수 · 24

1.3.4 다차원척도 그림의 해석 · 26

1.4 다차원척도법을 위한 SAS/PROC MDS의 예 · 27

1.4.1 [사례 1]을 위한 프로그램 및 결과 · 27

1.4.2 [사례 2]를 위한 프로그램 및 결과 · 29

1.4.3 [사례 3]을 위한 프로그램 및 결과 · 32

연습문제 · 34

2장 비유사성과 유사성

2.1 비유사성 · 40
2.1.1 비유사성 측도 · 40
2.1.2 비유사성 측도의 적용 · 43
2.1.3 비유사성 측도의 계산을 위한 프로그램 · 44
2.2 유사성 · 48
2.2.1 유사성 측도 · 48
2.2.2 유사성 측도의 적용 · 51
2.3 비유사성과 유사성의 변환 · 53
2.4 비대칭 비유사성과 유사성의 변환 · 56
연습문제 · 63

3장 이원 다차원척도법

3.1 계량형 다차원척도법 · 68
3.1.1 토거선 알고리즘 · 68
3.1.2 형상공간의 적합도 · 69
3.1.3 계량형 다차원척도법의 예 · 70
3.2 비계량형 다차원척도법 · 74
3.2.1 크루스칼-세퍼드 알고리즘 · 75
3.2.2 상그림과 세퍼드그림 · 77
3.2.3 비계량형 다차원척도법의 예 · 77
3.3 계량형과 비계량형 다차원척도법의 비교 · 83
3.3.1 SAS/MDS에서 비교 · 84
3.3.2 R-프로그램에서 비교 · 88
연습문제 · 91

4장 삼원 다차원척도법

4.1 삼원 다차원척도법을 위한 모형 · 100
4.1.1 INDSCAL 모형의 소개 · 100
4.1.2 INDSCAL 모형의 기하적 해석 · 103
4.2 삼원 다차원척도법의 예 · 109
4.2.1 신체 부위 자료 · 109
4.2.2 프로그램과 결과 · 110
4.2.3 결과의 해석 · 113
연습문제 · 116

5장 최대우도 다차원척도법

5.1 최대우도 다차원척도법을 위한 모형 · 120
5.1.1 램지모형의 소개 · 120
5.1.2 등분산 가정에서 최대우도추정 · 122
5.2 최대우도 다차원척도법의 예 · 125
5.2.1 레크리에이션 자료 · 125
5.2.2 프로그램과 결과 · 125
5.2.3 결과의 해석 · 129
5.3 삼원과 최대우도 다차원척도법의 비교 · 130
연습문제 · 135

6장 베이즈 다차원척도법

6.1 베이즈 계량형 다차원척도법 · 138

6.1.1 베이즈 모형 · 138

6.1.2 MCMC 알고리즘 · 141

6.1.3 차원의 수 · 143

6.2 베이즈 계량형 다차원척도법의 예 · 146

6.2.1 유럽 10대 도시 간의 거리 자료 · 146

6.2.2 R-프로그램과 결과 · 146

6.2.3 결과의 해석 · 149

6.3 계량형과 베이즈 계량형 다차원척도법의 비교 · 150

연습문제 · 151

7장 비선형 다차원척도법

7.1 비선형 계량형 다차원척도법 · 158

7.1.1 비선형 모형 · 158

7.1.2 비선형사상 알고리즘 · 159

7.2 비선형 계량형 다차원척도법의 예 · 161

7.2.1 모스부호 자료 · 161

7.2.2 R-프로그램과 결과 · 161

7.2.3 결과의 해석 · 163

7.3 계량형과 비선형 계량형 다차원척도법의 비교 · 164

연습문제 · 165

부록 1: SAS MDS 절차 · 167

부록 1.1 PROC MDS 문의 옵션 · 168

부록 1.1.1 측정수준 지정 · 169

부록 1.1.2 자료의 형태 · 170

부록 1.1.3 자료세트의 지정 · 171

부록 1.1.4 스트레스 및 최적척도화 · 172

부록 1.1.5 출력결과 조정 · 174

부록 1.1.6 수렴문제 조정 · 175

부록 1.1.7 결측치 조정 · 176

부록 1.1.8 기타 선택 · 176

부록 1.2 ODS GRAPHICS · 177

부록 1.3 PROC ALSCAL과 PROC MDS · 178

부록 1.4 PROC MLSCALE과 PROC MDS · 180

부록 2: 자료 · 182

부록 3: 프로그램 · 196

참고문헌 · 199

찾아보기 · 204

1장
다차원척도법의 이해

다차원척도법(MultiDimensional Scaling: MDS)에 대한 정의와 특히, 계량형(metric)과 비계량형(non-metric)인 이원(two-way)과, 삼원(three-way), 최대우도(maximum likelihood) 다차원척도법을 사례를 통해 미리 개괄적으로 소개하고 이해하려고 한다. 그리고 1950년대부터 많은 학자들에 의해서 개발되고 발전되어 온 다차원척도법의 역사적 배경을 살펴보고 측정수준, 스트레스, 차원의 수, 다차원척도그림(MDS map)의 해석을 중심으로 이론적 배경을 간단히 소개하려고 한다. 이 절에서 사용된 다차원척도법을 위한 프로그램인 SAS/PROC MDS의 문법과 본격적인 소개는 [부록 1: SAS MDS 절차]를 참고하기를 바란다.

1.1 다차원척도법이란

다차원척도법이란 다차원 공간에서 개체(objects) 간의 거리를 나타내는 자료로부터 그들의 유사성(similarity) 또는 비유사성(dissimilarity)을 저차원 공간에 기하적으로 나타내어 그들의 관계를 탐색적으로 살펴보는 다변량 그래프적 기법이다. 여기에서는 저차원 공간을 형상공간(configu ation space)이라고 하며, 여기에 개체를 기하적으로 나타낸 것을 형상 또는 다차원척도그림이라고 한다. 이 책에서는 후자인 다차원척도그림이라고 하자.

일반적으로 n개의 개체 간의 거리는 그들 간의 비유사성 d_{ij}, $i,j=1,...,n$를 측정한 것이다. 따라서 다차원척도법은 다음과 같은 n개의 개체 간의 비유사성을 나타내는 크기가 $n \times n$인 비유사성행렬(dissimilarity matrix)

$$D=(d_{ij}),\ \ i,j=1,...,n \tag{1.1}$$

을 구하고 이 비유사성행렬 D를 형상공간이라는 저차원 공간에 기하적으로 나타낸 다차원척도그림을 제공하는 기법이다. 여기서 일반적으로 비유사성행렬 D는 동일 개체의 비유사성을 나타내는 대각원소값들이 0인 대칭행렬이다.

다차원척도법은 비유사성행렬 D를 구성하는 개체 간의 비유사성 d_{ij}의 측정척도(measure)에 따라 계량형 다차원척도법과 비계량형 다차원척도법으로 나눌 수 있다(Kruskal과 Wish, 1978, pp. 19-23; Mardia 외 2인, 1979, Chapter 14; Johnson과 Wichern, 2002, pp. 700-708).

대개 계량형 다차원척도법에서는 비유사성 d_{ij}의 측정척도가 개체 간의 실제 측정거리 또는 매우 고전적인 유클리드거리(Euclidean distance)를 나타내며 비계량형 다차원척도법에서는 거리들의 크기 순서를 나타낸다. 특히, 허명회(1994, 8장)는 전자의 경우 절대척도(absolute scaling)와 구간척도(interval scaling)로, 후자의 경우는 순서척도(ordinal scaling)로 구분하고 있다. 이에 대해서는 1장 3절에서 자세히 다루기로 하고 유클리드거리 및 거리로 정의되는 비유사성의 개념 및 종류에 대해 2장에서 자세히 설명하고 있다.

이런 이유에서 계량형 다차원척도법을 고전적척도법(classical scaling), 비계량형 다차원척도법을 순서적척도법(ordinal scaling)으로 각각 언급하기도 한다(Chatfield와 Collins, 1980, Chapter 10; du Toit 외 2인, 1986, Chapter 6). 이들의 수리적 알고리즘은 엄격히 다르지만 행렬 D가 실제 측정값이나 유클리드거리의 자료인 경우 이들은 대동소이한 결과를 제공한다고 알려져 있다(Chatfield와 Collins, 1980, pp. 209-210; Mardia 외 2인,

1979, p. 415). 그러나 일반적으로 비계량형 다차원척도법이 계량형 다차원척도법보다는 대개 바람직한 결과를 제공한다고 알려져 있다(Kruskal과 Wish, 1978, pp. 76-78; Chatfield와 Collins, 1980, pp. 209-210).

광범위하게는 계량형과 비계량형 다차원척도법을 이원 다차원척도법이라고 하며 이는 이원행렬인 비유사성행렬 D에 적용된 다차원척도법을 뜻한다. 이는 3장에서 자세히 다루고 있다. 반면에 개체 간의 유클리드거리를 이용한 식 (1.1)에서 정의한 n개 개체 간의 비유사성 d_{ij} 대신에 가중(weighted)유클리드거리를 이용한 크기가 $n \times n$ 이원행렬 D를 m명의 주체(subjects) 각각에 대해서 고려한 삼원행렬

$$D^{(s)} = (d_{ij}^{(s)}), \quad i, j = 1, \ldots, n; \; s = 1, \ldots, m. \tag{1.2}$$

을 이용한다. 여기서, $d_{ij}^{(s)}$는 s번째 주체에 대해 i번째와 j번째 개체 사이의 가중유클리드거리에 의한 비유사성을 나타낸다. Kruskal과 Wish(1978, p. 60), du Toit 외 2인(1986, pp. 144-151), 그리고 Everitt와 Dunn(1991, pp. 82-88)에 따르면 식 (1.2)에 대한 다차원척도법을 삼원 다차원척도법이라고 하며 가중 다차원척도법이라고도 한다. 이를 위한 알고리즘으로 Carroll과 Chang(1970)이 제안한 INDSCAL(individual differences scaling)을 수정 보완하여 교대최소제곱(alternatives least squares)법을 사용하는 Takane 외 2인(1977)의 ALSCAL을 채택하고 있으며 수치해석적으로 뉴턴-랩슨(Newton-Raphson) 방법에 의해서 계산되며 4장에서 다루고 있다.

5장에서는 4장의 삼원 다차원척도법에서 고려하지 않은 측정오차(measure error)에 대한 분포를 가정하고 최대우도추정법(maximum likelihood estimation method)을 활용한 최대우도 다차원척도법을 소개하고 있다. 이는 Ramsay(1977, 1980, 1982)의 알고리즘으로 수치해석적으로 비선형 최적화(nonlinear optimization)를 따른다.

6장에서는 5장의 최대우도추정법 대신에 베이즈(Bayesian) 추정과 MCMC 알고리즘을 적용한 Oh와 Raftery(2001)의 베이즈 다차원척도법(Bayesian MDS)을 정리하고 있다. 이를 바탕으로 다양한 상황에서 Okada와 Shigemasu(2009, 2010), Okada와 Mayekawa(2011) 그리고 Okada(2012)도 베이즈 추정법을 적용한 다양한 베이즈 다차원척도법을 소개하고 있다. 특히, Okada와 Shigemasu(2009)는 베이즈 다차원척도법의 알고리즘을 위한 Oh와 Raftery(2001)의 FORTRAN 프로그램을 R-프로그램으로 전환하여 제공하고 있다. 특히, Izenman(2008, Chapter 13)은 계량형 다차원척도법의 한 방법으로 베이즈 다차원척도법을 소개하고 있다.

7장에서는 Sammon(1969)의 비선형사상 알고리즘을 적용한 새먼사상(Sammon mapping)을 다차원척도법에 적용한 것을 비선형 다차원척도법(nonlinear MDS)이라고 하자. 근본적으로 비선형 다차원척도법은 계량형 다차원척도법의 스트레스 최소화 알고리즘의 비선형 버전이며 sammon 함수를 활용하는 R-프로그램에서 제공된다.

다차원척도법을 위한 프로그램으로 가장 활용도가 높은 SAS/PROC MDS는 SAS/STAT의 6.07판 이후에 현재 9.2판인 SAS Institute Inc.(2008, Chapter 53)에 나와 있고 이 책에서는 이를 응용의 관점에서 주로 활용하고자 한다. 더불어 행렬언어를 이용하는 SAS/IML과 R도 필요에 따라서 활용되고 있다.

1.1.1 [사례 1] 우리나라 도시 간 철도거리

철도청(1993)의 철도통계연보 부록의 철도영업키로정표로부터 우리나라 8대 도시 간 최단 철도거리에 대한 자료가 〈표 1.1〉에 있다. 여기에서 이 거리를 비유사성 $d_{ij}, \quad i,j=1,\ldots,8$으로 정의하면 이는 실제 측정거리로 비유사성행렬 D를 구성한다. 이 경우 일반적으로 계량형 다차원척도법을 적용하며 〈그림 1.1〉의 2차원 다차원척도그림을 얻을 수 있다. 이를 위한 SAS 프로그램과 실행결과는 1.4.1절에 정리되어 있다.

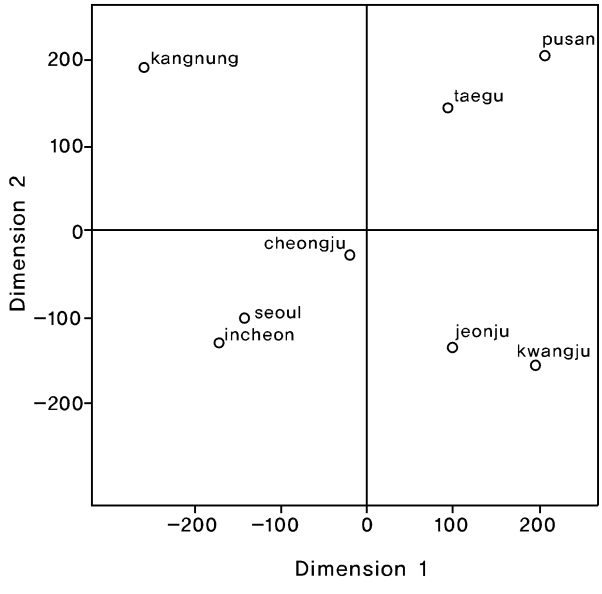
〈표 1.1〉 도시 간 철도거리

(단위: km)

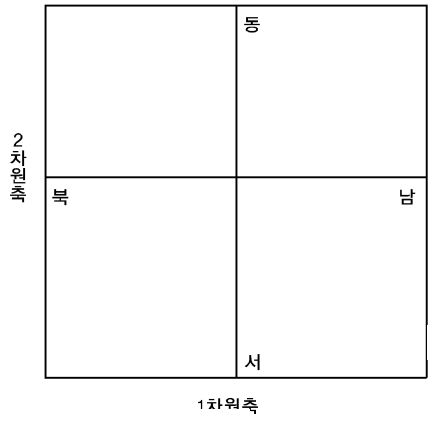
도시	서울	부산	광주	대구	인천	강릉	청주	전주
서울	0.0							
부산	444.5	0.0						
광주	353.8	364.7	0.0					
대구	323.9	120.6	350.0	0.0				
인천	30.9	475.4	384.7	359.9	0.0			
강릉	348.5	505.0	523.0	388.8	379.4	0.0		
청주	141.2	326.1	235.4	211.4	172.1	288.6	0.0	
전주	279.9	391.0	124.6	276.4	310.8	449.4	136.3	0.0

〈그림 1.1〉-(a)의 다차원척도그림을 원점을 중심으로 시계 방향으로 90도 정도 회전한다면 도시들이 실제 우리나라 지도상에 위치하고 있는 것과 같다. 〈그림 1.1〉의 수평축은 남-북으로 도시들을 분할하고 있고 수직축은 동-서로 도시들을 분할하고 있다. 일반적으로 〈그림 1.1〉-(a)에 대한 해석도 (b)를 보면 수평축을 1차원 축(Dimension 1)이라 하고 수직축을 2차원 축(Dimension 2)이라고 한다. 따라서 1차원 축을 남-북 방향, 2차원 축을 동-서

방향의 축으로 해석할 수 있다. 이와 같이 지리적 거리에 대한 사례들은 다차원척도법을 소개할 때 대개 많은 문헌의 앞머리에서 언급된다.



(a) 계량형 다차원척도그림



(b) 해석도

〈그림 1-1〉 도시 간 철도거리의 계량형 다차원척도그림과 해석도

1.1.2 [사례 2] 우리나라 경제관련기관의 경제 전망

〈표 1.2〉는 우리나라 14개 경제관련기관의 전문가들의 향후 1년간 10가지 경제 전망에 관한 조사표(동아일보, 1989. 1. 4)인 계량치 자료이며 장광국(1990, p. 13)이 잘 요약하고 있다. 이 자료의 유클리드거리와 이 자료로부터 변환된 이진수 자료(binary data)의 제곱 유클리드거리에 의한 두 가지 비유사성행렬을 제공하고 이에 대한 비계량형 다차원척도법을 실시하고 비교하려고 한다.

〈표 1.2〉 우리나라 14개 경제관련기관의 향후 1년간 10가지 경제 전망

기관	성장률	GNP	수출	수입	국제 흑자	연말 외채	연말 환율	실업률	소비 물가	임금 상승
한국은행	8.2	1950	698	650	98	280	630	3.0	5.7	12.0
대우증권	9.5	2100	710	620	110	270	640	2.7	4.5	12.0
동서증권	9.0	2000	690	630	100	290	630	2.6	6.0	12.0
전경련	7.8	1850	668	660	88	280	620	3.0	6.4	13.8
대한상공회의소	8.5	1928	710	670	90	290	620	3.0	6.0	10.0
중소기업중앙회	9.0	1958	710	615	95	280	603	4.0	6.0	10.0
현대자동차	8.5	1900	700	610	100	250	620	3.2	5.5	14.0
삼성물산	8.0	1900	700	640	100	280	640	2.7	5.5	10.0
선경	8.5	1950	700	620	120	300	630	2.7	7.0	12.0
대우경제연구소	7.9	1900	697	645	95	280	610	2.9	6.8	15.0
신한종합연구소	9.0	2030	700	620	100	275	630	3.0	7.0	13.0
동서경제연구소	8.5	1950	690	630	90	290	630	2.9	6.0	12.0
고려대학교	9.9	1870	729	649	123	260	616	3.4	6.1	15.8
중앙대학교	8.5	1700	700	630	90	260	620	4.0	6.0	14.0

먼저 〈표 1.2〉로부터 다차원척도법을 적용하기 위한 비유사성행렬 D를 구성할 비유사성 $d_{ij,}\ i,j=1,\ldots,14$를 1장 3절의 식 (1.3)에서 정의한 유클리드거리로 계산한 것이 〈표 1.3〉이다. 물론 10가지 경제 전망의 측정 단위가 다르므로 이들을 표준화하였다.

〈표 1.3〉 유클리드거리에 의한 비유사성행렬

한국은행	0.000													
대우증권	4.166	0.000												
동서증권	2.293	3.672	0.000											
전경련	3.269	6.925	4.110	0.000										
대한상공회의소	2.390	5.476	3.496	4.157	0.000									
중소기업중앙회	4.410	5.723	4.635	5.837	4.316	0.000								
현대자동차	3.573	4.432	4.084	4.812	5.177	4.231	0.000							
삼성물산	1.846	4.197	2.839	4.322	3.182	5.220	4.233	0.000						
선경	3.700	5.189	2.854	5.207	4.631	5.231	5.088	3.880	0.000					
대우경제연구소	3.131	6.345	3.800	2.789	3.820	4.796	3.928	4.468	4.159	0.000				
신한종합연구소	3.118	4.344	2.363	4.653	4.265	4.260	3.592	3.920	2.990	3.549	0.000			
동서경제연구소	1.784	4.443	1.515	3.291	3.061	4.263	3.820	2.469	3.401	3.278	2.648	0.000		
고려대학교	5.362	5.571	5.684	6.926	5.856	5.563	4.699	6.336	5.696	5.349	4.994	6.088	0.000	
중앙대학교	4.343	6.693	5.425	4.474	5.124	4.407	3.336	5.114	6.064	4.192	4.934	4.568	5.176	0.000

다음으로 허명회(1992, 4장)가 〈표 1.2〉의 자료에서 10가지 경제 전망 항목을 주관적 기준으로 예(1)와 아니오(0) 형태의 이진수 자료로 가공하여 〈표 1.4〉를 제공하였다. 여기서 10가지 경제 전망 항목의 주관적 기준은 다음과 같다.

(1) 성장률이 8.5% 이상이 될 것이다.
(2) GNP는 1950달러 이상이 될 것이다.
(3) 수출은 700억 달러 이상이 될 것이다.
(4) 수입은 630억 달러 이하가 될 것이다.
(5) 국제 흑자는 100억 달러 이하가 될 것이다.
(6) 연말 외채는 280억 달러 이하가 될 것이다.
(7) 연말 환율은 1달러당 630원 이상일 것이다.
(8) 실업률은 3.0% 이하일 것이다.
(9) 소비자 물가의 상승률은 6.0% 이하일 것이다.
(10) 임금 상승률은 12.5% 이하일 것이다.

〈표 1.4〉 경제관련기관의 10가지 경제 전망의 이진수 자료

기관	경제 전망									
	1	2	3	4	5	6	7	8	9	10
한국은행	0	1	0	0	0	1	1	1	1	1
대우증권	1	1	1	1	1	1	1	1	1	1
동서증권	1	1	0	1	1	0	1	1	1	1
전경련	0	0	0	0	0	1	0	1	0	0
대한상공회의소	1	0	1	0	0	0	0	1	1	1
중소기업중앙회	1	1	1	1	0	1	0	0	1	1
현대자동차	1	0	1	1	1	1	0	0	1	0
삼성물산	0	0	1	0	1	1	1	1	1	1
선경	1	1	1	1	1	0	1	1	0	1
대우경제연구소	0	0	0	0	0	1	0	1	0	0
신한종합연구소	1	1	1	1	1	1	1	1	0	0
동서경제연구소	1	1	0	1	0	0	1	1	1	1
고려대학교	1	0	1	0	1	1	0	0	0	0
중앙대학교	1	0	1	1	0	1	0	0	1	0

이진수 자료 〈표 1.4〉로부터 다차원척도법을 적용하기 위한 비유사성 $d_{ij,}\ i,j=1,\ldots,14$를 다음과 같이 정의하자.

$$d_{ij}=\frac{i\text{번째와 } j\text{번째 개체간에 일치하지 않는 항목의 수}}{\text{전체 항목의 수}}.$$

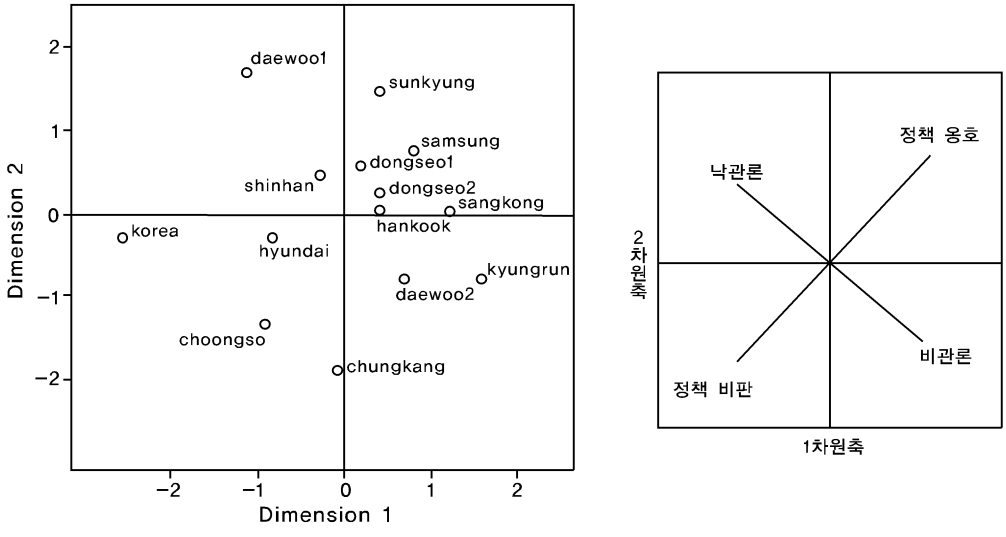
예를 들면 한국은행과 대우증권 두 기관은 10가지 경제 전망 항목 중 4가지가 일치하지 않는다. 따라서 그들의 비유사성은 4/10 = 0.4이므로 〈표 1.5〉와 같은 비유사성행렬 자료를 얻을 수 있다. 이는 2.1.2절에서 보다 자세히 다시 언급될 것이며, 특히 이진수 자료로부터 개체 i와 j의 유사성 c_{ij}를 계산하는 방법 중에서 단순매칭계수(simple matching coefficient)를 이용하여 비유사성 $d_{ij}=1-c_{ij}$를 구하는 것과 연관성이 있다. 2.1.2절을 보면 유사성의 수치적 계산은 이진수 자료의 제곱유클리드거리에 의해서 이루어진다.

〈그림 1.2〉는 〈표 1.5〉의 2차원 다차원척도그림이다. 이를 위한 SAS 프로그램과 실행결과는 1.4.2절을 참고하기를 바란다.

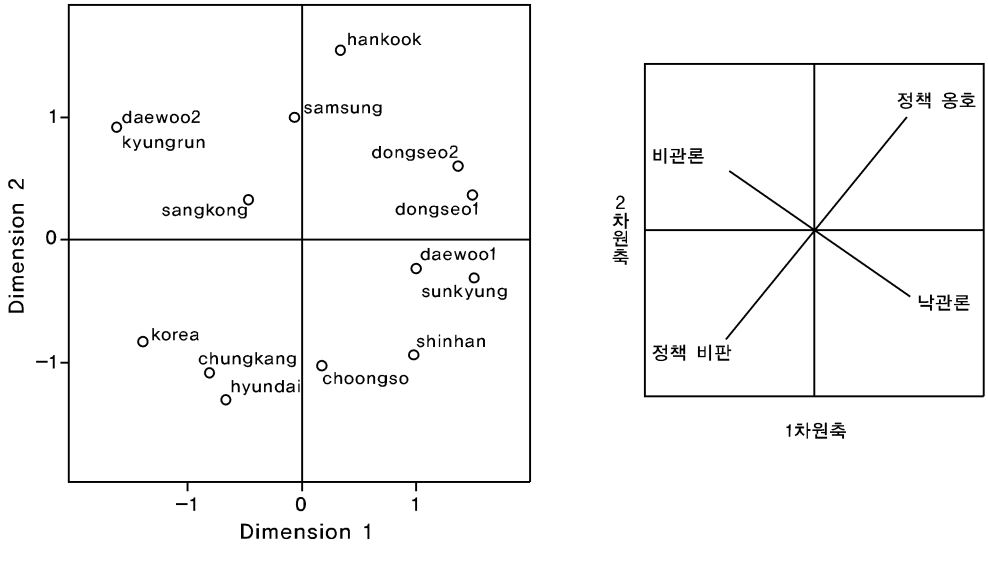
〈표 1.5〉 이진수 자료의 제곱유클리드거리에 의한 비유사성행렬 자료

한국은행	.0													
대우증권	.4	.0												
동서증권	.4	.2	.0											
전경련	.4	.8	.8	.0										
상공회의소	.5	.5	.5	.5	.0									
중소기업중앙회	.5	.3	.5	.7	.4	.0								
현대자동차	.8	.4	.6	.6	.5	.3	.0							
삼성물산	.3	.3	.5	.5	.4	.6	.5	.0						
선경	.6	.2	.2	.8	.5	.5	.6	.5	.0					
대우경제연구소	.4	.8	.8	.0	.5	.7	.6	.5	.8	.0				
신한종합연구소	.6	.2	.4	.6	.7	.5	.4	.5	.2	.6	.0			
동서경제연구소	.3	.3	.1	.7	.4	.4	.7	.6	.3	.7	.5	.0		
고려대학교	.8	.6	.8	.4	.5	.5	.2	.5	.6	.4	.4	.9	.0	
중앙대학교	.7	.5	.7	.5	.4	.2	.1	.6	.7	.5	.5	.6	.3	.0

〈그림 1.2〉의 (a)와 (b)에서 나타나는 기관들의 경제 전망은 입력 자료가 다르지만 서로 유사하며 약간의 차이점이 있는데, 이는 해석도를 참고로 다음과 같이 요약될 수 있다. 먼저 경제 전망을 연결한 선은 새로운 축으로 원차원 축을 회전시킨 것과 같다. (a)와 (b)에서 비관론-낙관론 차원축의 방향은 다르지만 공통적으로 kyungrun(전경련)과 daewoo2(대우경제연구소)가 비관론 쪽에 있으며 반대편인 낙관론 쪽에 daewoo1(대우증권), shinhan(신한종합연구소)이 있다. 특히, (a)의 정책옹호 쪽에 있는 sunkyung(선경), samsung(삼성물산), sangkong(상공회의소) 중 sunkyung(선경)은 (b)의 낙관론으로, samsung(삼성물산)과 sangkong(상공회의소)는 (b)의 비관론으로 분류되어 있다. sunkyung(선경)이 (a)에서 나머지 기관들과 다소 떨어져 있고 해석도와 같이 새로운 축에 대해서 낙관론 쪽으로 분류될 수는 있으나, samsung(삼성물산)이 (a)의 정책옹호에서 (b)의 비관론으로 분류되는 것은 (a)와 (b)의 해석상 상당한 차이를 보이는 점이다. 정책비판 축에 대해서는 (a)의 korea(고려대학교), hyundai(현대자동차), choongso(중소기업중앙회), chungang(중앙대학교)가 위치하고 있으나 이들 중 choongso(중소기업중앙회)가 (b)에서는 정책비판-낙관론 두 축의 중앙에 위치하고 있어 중립적 경제 전망을 하고 있다. 이상에서 몇 개의 경제관련 기관의 전망이 다차원척도그림 (a)와 (b)에 따라 차이점을 보이는 것은 (b)의 경우 이진수 자료를 만드는 과정에서 분석자의 주관적 기준이 개입되어 있기 때문이다. 그러나 전체적으로 (a)와 (b)의 다차원척도법은 유사한 결과와 해석을 제공하고 있다.



(a) 원 자료의 유클리드거리



(b) 이진수 자료의 제곱유클리드거리

〈그림 1-2〉 경제관련기관의 경제 전망의 비계량형 다차원척도그림과 해석도

1.1.3 [사례 3] 12개 국가의 유사성에 대한 평가

18명의 학생들이 12개 국가를 각각 두 나라씩 짝지은 66가지 경우에 대해 1점(매우 다르다)~9점(매우 유사하다)의 9점 척도로 응답하여 그 평균으로 12개국의 유사성행렬 〈표 1.6〉을 얻었다(Kruskal과 Wish, 1978, pp. 30-31; Everitt와 Dunn, 1991).

예) Brazil과 Congo

매우 유사하다(9점) _____ 매우 다르다(1점)
(very similar) (very different)

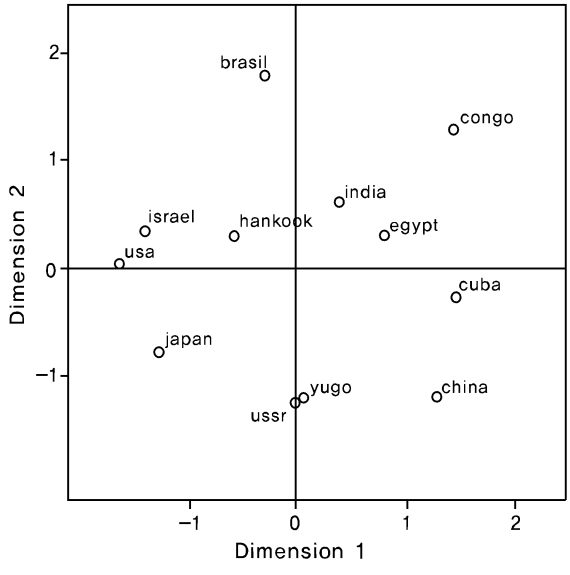
〈표 1.6〉 12개 국가의 유사성행렬 자료

Brazil	.											
Congo	4.83	.										
Cuba	5.28	4.26	.									
Egypt	3.44	5.00	5.17	.								
France	4.72	4.00	4.11	4.78	.							
India	4.50	4.83	4.00	5.83	3.44	.						
Israel	3.83	3.33	3.61	4.67	4.00	4.11	.					
Japan	3.50	3.39	2.94	3.84	4.11	4.50	4.83	.				
China	2.39	4.00	5.50	4.39	3.67	4.11	3.00	4.17	.			
USSR	3.06	3.39	5.44	4.39	5.06	4.50	4.17	4.61	5.72	.		
USA	5.39	2.39	3.17	3.33	5.94	4.28	5.94	6.06	2.56	5.00	.	
Yugoslav	3.17	3.50	5.11	4.28	4.72	4.00	4.44	4.28	5.06	6.67	3.56	.

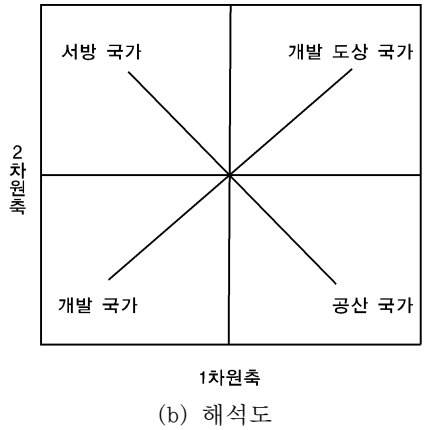
이 자료는 앞서 살펴본 비유사성행렬 자료들과는 달리 유사성행렬 자료이다. 여기서는 개체 i와 j 간의 유사성을 c_{ij}라고 하면 자기 자신의 유사성은 9점(매우 유사하다)이므로 변환 $(9-c_{ij})=d_{ij}$를 고려하면 비유사성을 얻을 수 있다. 유사성행렬 자료로부터 비유사성행렬 자료로 바꾸는 변환 방법은 2.3절에서 자세히 설명하고 있다.

〈그림 1.3〉은 〈표 1.6〉의 유사성행렬 자료에 대한 2차원 비계량형 다차원척도그림과 해석도이다. 이를 위한 SAS 프로그램과 실행결과는 1.4.3절의 [프로그램 1.3]과 [결과 1.3]을 참고하기를 바란다. 〈그림 1.3〉의 해석도에서 새롭게 연결된 축은 정치적노선(서방국가-공

산국가)과 경제개발(개발국가-개발도상국가) 차원축으로 해석할 수 있다. 보다 자세한 설명은 Kruskal과 Wish(1978, pp. 30-43)를 참고하기를 바란다.



(a) 비계량형 다차원척도그림



(b) 해석도

〈그림 1-3〉 12개 국가의 유사성에 대한 비계량형 다차원척도그림과 해석도

1.1.4 [사례 4] 15군데 신체 부위의 비유사성

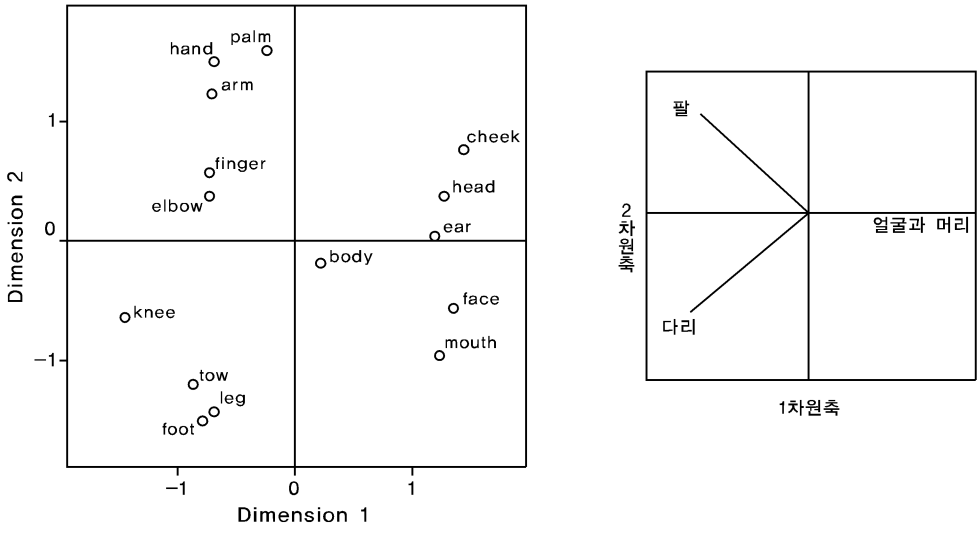
〈표 1.7〉은 Young과 Lewyckyj(1979, pp. 105–108)의 아동과 성인에게 15군데 신체(cheek, face, mouth, head, ear, body, arm, elbow, hand, palm, finger, leg, knee, foot, toe)의 비유사성을 서로 판단하도록 조사한 자료의 일부이다. 이는 신체 부위에 대한 아동이 판단하는 특성과 이러한 판단들이 성인이 됨에 따라 변화하는 방향을 파악하는 데 의미가 있다. 이 조사에는 나이가 6세, 8세, 10세의 아동들 3그룹과 대학교 2년생을 1그룹으로 하여 총 4그룹을 구성하였고 각 그룹당 15명씩 배당되어 총 60명의 주체(subjects)에 대해서 이루어졌다.

〈표 1.7〉 15군데 신체 부위에 대한 비유사성 자료(일부)

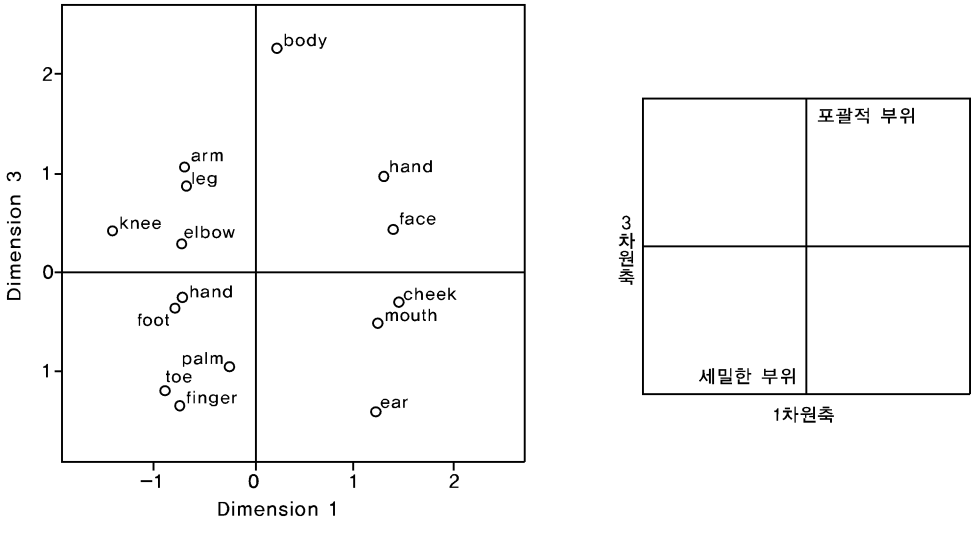
subject	cheek	face	mouth	head	ear	body	arm	elbow	hand	palm	finger	leg	knee	foot	toe
1	0	2	1	3	4	10	5	9	6	7	8	11	12	13	14
1	2	0	12	1	13	3	8	10	11	9	7	4	5	6	14
1	3	2	0	1	4	9	5	11	6	7	8	10	13	12	14
1	2	1	3	0	4	9	5	6	11	7	8	10	12	13	14
1	10	1	11	2	0	6	3	4	5	12	13	7	8	14	9
1	14	12	9	6	13	0	8	7	5	10	11	1	4	2	3
1	12	14	11	10	13	5	0	4	1	3	2	6	9	7	8
1	5	7	14	8	6	9	1	0	2	3	4	10	11	12	13
1	13	11	12	10	14	9	3	4	0	1	2	6	5	7	8
1	8	6	7	9	4	5	3	10	1	0	2	12	11	13	14
1	4	5	13	6	9	12	3	4	1	2	0	7	8	10	11
1	14	12	13	11	9	7	4	6	5	3	10	0	8	1	2
1	12	11	14	10	13	4	5	8	6	7	9	1	0	2	3
1	12	14	10	13	11	9	4	5	8	6	7	2	3	0	1
1	13	8	9	11	14	3	6	5	7	10	12	2	4	1	0

입력 자료인 비유사성행렬 D를 구성하는 〈표 1.7〉을 살펴보면 이는 15×15인 정방행렬로 구성되어 있다. 이는 주체인 한 사람이 15군데 신체 부위를 각각 돌아가면서 표준부위로 두고 15점 척도로 이와 나머지 신체 위와의 유사성을 판단한 순위–순서형(rank–order) 자료이다. 다시 말해서 첫 번째 행은 첫 번째 신체 부위인 cheek을 표준개체(부위)인 0점으로 두고 나머지 14개 신체 부위에서 이와 가장 유사한 개체인 mouth를 1점으로 하고, 두 번째 유사한 개체인 face를 2점으로 하여 14점을 받은 toe는 매우 다른 개체의 순위를 말한다. 두 번째 행은 두 번째 신체 부위인 face를 표준개체 0점으로 두고 첫 번째 행처럼

나머지 부위와의 유사성에 따라서 순위를 매겨간다. 따라서 이렇게 만들어진 비유사성행렬 D의 대각원소는 항상 0이고, 비대각원소는 비대칭으로 이루어져 있다는 점이다. [부록 2: 자료 1(파일이름: body.txt)]에는 분석 편의상 6세 아동 15명과 대학교 2년생 15명으로 이루어진 두 그룹에 대한 총 30명에 대해 〈표 1.7〉과 같은 15×15 비유사성행렬 D가 30개



(a) 1차원축과 2차원축



(b) 1차원축과 3차원축

〈그림 1.4〉 15군데 신체 부위에 대한 삼원 다차원척도그림과 해석도

가 정리되어 있다. 즉, 식 (1.2)의 표현에 따르면 30명의 주체($s=1,\ldots,30$) 각각에 대하여 비유사성행렬 $D^{(s)}$이 존재하여 전체적으로 삼원행렬로 이루어진다. 실제로 〈표 1.7〉은 $D^{(1)}$에 해당한다.

이런 형태의 자료에 대해서는 삼원 다차원척도법을 적용하며 4장에서 [프로그램 4.2]와 [결과 4.2]에 잘 정리되어 있다. 여기에서는 [결과 4.2]의 삼원 다차원척도그림을 먼저 인용한 〈그림 1.4〉의 (a)와 (b)를 통해 삼원 다차원척도법을 이해하려고 한다.

15개 신체 각 부위에 대한 삼원 다차원척도그림은 차원별로 조합되어 주어져 있다. 먼저 1차원축(Dimension 1)과 2차원축(Dimension 2)에 의한 〈그림 1.4〉-(a)는 각 신체 부위가 각 정점에 놓인 삼각형 모양을 나타내고 있다. 특히, 정점의 세 신체 부위인 arm(팔), leg(다리), head(머리)는 각 군집으로 중심의 body(몸)와 구분되고 있다. 구체적으로 수평축인 1차원축은 왼편의 팔과 다리 부위와 오른편의 머리 부위로 분할하고 있고 수직축인 2차원축은 위쪽 팔 부위와 아래쪽 다리 부위로 분할하고 있다. 〈그림 1.4〉-(b)에서는 3차원축(Dimension 3)의 위쪽으로는 body(몸), head(머리), face(얼굴), 다리(leg), arm(팔) 등과 같이 신체 중 포괄적으로 큰 부위를 나타내고 아래쪽으로는 mouth(입), ear(귀), hand(손), palm(손바닥), finger(손가락), toe(발톱) 등과 같이 큰 부위의 세밀한 부위를 나타내고 있다.

1.1.5 [사례 5] 15가지 레크리에이션의 비유사성

Ramsay(1983)는 10명에게 15가지 레크리에이션(concert, museum, theatre, movie, watch TV, conference, reading, watch hockey, ballet, political debate, fashion show, documentary film, exhibition, window shopping, restaurant)의 비유사성을 판단하기 위해 조사를 수행하였다. 이 조사는 레크리에이션을 각각 둘씩 짝지은 105가지 경우에 대해 다음과 같이 비유사성의 정도를 25점 척도로 표시하도록 한 설문지로부터 얻어진 것이다.

예) Museum과 Hockey

매우 유사하다(0점)	_____	매우 다르다(24점)
(very similar)		(very different)

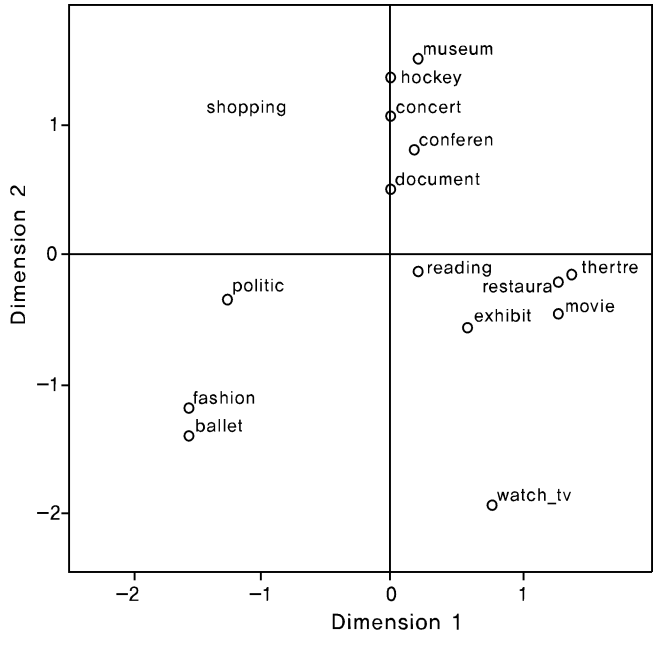
완전한 Ramsay의 자료는 [부록 2: 자료 2(파일이름: recreation.txt)]에 정리해 두었다. 이로부터 〈표 1.8〉은 사례에서 자료를 이해하고자 일부를 정리한 것으로 첫 번째 설문 응답자의 자료이다.

따라서 식 (1.2)에 따르면 〈표 1.8〉은 크기가 15×15인 비유사성행렬 D에 해당하며 설문 응답자 10명의 주체에 대하여 이와 같은 행렬 10개가 존재하는 삼원행렬자료 $D^{(s)}$, $s = 1, \ldots, 10$중 $D^{(1)}$에 해당한다. 이 자료에 대해 본 사례에서는 최대우도 다차원척도법을 적용하였고 이는 광범위하게는 1.1.4절의 [사례 4]와 같이 4장의 삼원 다차원척도법으로 분류할 수 있다. 최대우도 다차원척도법의 알고리즘과 관련된 SAS 프로그램은 5장에 정리해 두었으니 참고하기를 바란다.

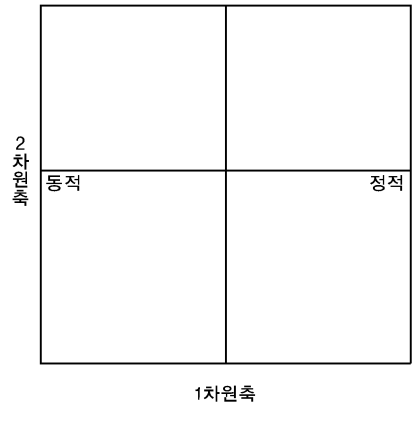
〈표 1.8〉 15가지 레크리에이션의 비유사성자료(일부)

concert	0														
museum	8	0													
theatre	18	21	0												
movie	5	22	16	0											
watch TV	8	24	22	11	0										
conference	11	13	19	18	22	0									
reading	19	12	15	9	21	14	0								
watch hockey	6	6	15	19	21	13	23	0							
ballet	24	18	21	15	9	19	16	23	0						
political debate	21	21	23	23	23	13	12	11	23	0					
fashion show	22	20	24	21	23	10	13	12	7	7	0				
documentary film	10	7	21	11	9	10	6	8	14	22	22	0			
exhibition	23	21	18	9	8	10	7	14	22	9	11	8	0		
window shopping	5	14	19	16	23	22	19	20	8	17	23	21	20	0	
restaurant	14	23	11	8	21	22	7	10	23	19	21	9	7	22	0

〈그림 1.5〉는 레크리에이션의 비유사성에 대한 최대우도 다차원척도그림으로 5장의 [프로그램 5.1]에 의해서 얻을 수 있다. 특히, 그림에서 수평축인 1차원축(Dimension 1)은 왼쪽의 쇼핑(shopping), 정치적 논쟁(politic), 패션쇼우(fashion), 발레(ballet) 등과 같이 현장감 있고 동적인 레크리에이션을 나타내며, 오른쪽의 독서(reading), 영화(movie), 박물관(museum), 전시회(exhibit) 등은 주로 정적인 레크리에이션으로 구분하고 있다.



(a) 최대우도 다차원척도그림



(b) 해석도

〈그림 1.5〉 레크리에이션에 대한 최대우도 다차원척도그림과 해석도

1.2 다차원척도법의 역사적 배경

Young(1987, Chapter 2)은 다차원척도법의 역사적 배경으로 1950년에서 1980년까지의 30년간을 대략 10년 주기로 다음과 같이 요약 언급하고 있다. 다차원척도법의 개발의 첫 출발은 1950년대였다. Torgerson(1952)이 처음 다차원척도법을 제안하였고 Messick과 Abelson(1956)에 의해서 일반화되었다. 이들은 개체 간의 유클리드거리를 사용하였다. Attneave(1950)은 일찍이 이 거리 대신에 시티-블럭거리(city-block distance)를 제안하였지만 Torgerson처럼 저차원 공간에 개체 간의 관계를 나타내지는 못하였다. 다차원척도법에서 이와 같은 거리를 이용한 접근법은 계량형 다차원척도법으로 알려지게 되었다. 이들 거리에 대한 개념과 정의는 2장에서 언급하고 있다.

1960년대는 매우 활발히 다차원척도법 개발이 이루어진 때이다. 거의 모든 주된 연구는 순서화된 비유사성행렬자료의 분석을 위한 방법과 개발이었다. 이러한 방법들을 다차원척도법에서는 비계량형 다차원척도법이라 하였다. Shepard(1962)는 비계량형 다차원척도법에 대해 처음 소개하였고 그의 논문으로 인하여 이 시기에 다차원척도법에 대한 많은 연구가 있게 된 계기가 되었다. 그 중에서 오늘날 보편적으로 이용되고 있는 Kruskal(1964)의 방법은 개체를 나타낼 기하적 공간의 차원을 Shepard보다 더 축소할 수 있는 것이었다. Kruskal이 Shepard 방법을 수정 보완하여 오늘날 비계량형 다차원척도법의 알고리즘을 제시하였고, 이는 크루스칼-세퍼드 알고리즘이라고 불리고 있다(Mardia 외 2인, 1979, p. 414; Jobson, 1992, pp. 585-586). 이들 외에는 Torgerson과 Meuser(1962)과 Guttman (1968) 등의 다양한 방법들이 제시되기도 하였다.

지금까지(1950~1960년대)는 오직 하나의 크기가 $n \times n$인 비유사성행렬 D에 대한 다차원척도법을 논의하였다. 실제로 1장 1절의 [사례 1]-[사례 3]에서 살펴본 비유사성행렬 D를 이원행렬이라 하고 이에 적용된 다차원척도법을 이원 다차원척도법이라고 한다. 반면에 [사례 4]는 총 60명의 주체가 15명씩 4그룹으로 15군데의 신체 부위의 유사성을 판단하여 크기가 15×15인 비유사성행렬을 제공하였는데, 특히 본 사례에서는 편의상 두 그룹의 30명에 대한 비유사성행렬 $D^{(s)}$, $s = 1,..30$에 대한 분석을 제공하고 있다. [사례 5]에서는 설문 응답자 10명으로부터 15가지 레크리에이션에 대하여 크기가 15×15인 비유사성행렬 $D^{(s)}$, $s = 1,..10$을 구성하여 분석하였다. 이들 두 사례에서는 공통적으로 주체에 대하여 비유사성행렬이 존재하며 이를 삼원행렬(three-way matrix)이라고 한다. 그리고 이에 적용된 다차원척도법을 삼원 다차원척도법이라고 한다(Kruskal과 Wish, 1978, p. 60; Everitt와 Dunn, 1991, pp. 82-88). 1970년대는 이와 같은 삼원 다차원척도법을 위한 여

러 가지 알고리즘 개발(Tucker과 Messick, 1963; Carroll과 Chang, 1970)이 이루어진 때이다. 특히 Carroll과 Chang(1970)의 INDSCAL(individual differences scaling)은 각 주체에 대한 가중치를 고려한 가중(weighted)유클리드거리를 비유사성으로 하는 것이다. Takane 외 2인(1977)은 이 알고리즘을 수정하고 보완하여 ALSCAL(alternating least squares scaling)을 제시하였다. 이는 범용 SAS에서는 PROC ALSCAL로 범용 SPSS^{X}에서는 ALSCAL 절차로 각각 채택되어 활용되고 있는데, 삼원 다차원척도법은 물론이고 이원 다차원척도법까지 가능하게끔 되어 있다(Young, 1987, p. 36). PC-윈도우즈용 SAS/STAT 6.07판에서는 범용의 PROC ALSCAL의 많은 점을 수용하여 PROC MDS를 제공하였고, 9.2판인 SAS Institute Inc.(2008, Chapter 53)에서는 ODS GRAPHICS 문을 활용하여 그래픽 기능을 강화하고 사용자가 편하게 사용하도록 하였다.

1980년대 이후 Ramsay(1982)의 최대우도(maximum likelihood) 다차원척도법이 개발되어 범용 SAS에서 PROC MLSCALE로 제공되었다. 윈도우즈용 SAS의 PROC MDS에서 이와 유사한 몇 가지 결과를 제공할 수 있다. 자세한 설명은 5장의 최대우도 다차원척도법으로 다루기로 한다. 2000년대에는 베이즈 추정과 MCMC 알고리즘을 적용한 Oh와 Raftery (2001)의 베이즈 다차원척도법(Bayesian MDS)이 개발되었고, 특히 Izenman(2008, 13장)은 계량형 다차원척도법의 한 방법으로 베이즈 다차원척도법을 소개하고 있다. 더불어 패턴인식(pattern recognition)에서 잘 알려진 도구인 Sammon(1969)의 비선형사상 알고리즘을 적용한 비선형 다차원척도법(nonlinear MDS)도 Dzwinel(1994), Johnson과 Wichern (2002, p. 708) 그리고 Izenman(2008, pp. 488-490)에서 찾아볼 수 있다. 추가적으로 자료에 이상치가 있는 경우 이들의 영향을 줄이는 알고리즘을 적용한 로버스트(robust) 다차원척도법에 대한 연구와 활용은 Choi 외 2인(1998), Cox와 Cox(2001, pp. 96-98) 그리고 Forero와 Giannakis(2012)에서 살펴볼 수 있다.

다차원척도법은 정치학(이재창과 박정섭, 1986; Young, 1987, Chapter 10), 경제학(허명회, 1994, 8장), 심리학(Young, 1987, Chapter 8; Kruskal과 Wish, 1984, pp. 30-35; Chatfield와 Collins, 1980, pp. 194-195; Boyle과 Katz, 1991, pp. 565-574), 사회학(du Toit 외 2인, 1986, pp. 127-131), 시장조사(이재창과 박정섭, 1986; Young, 1987, Chapter 9; Lawless 외 2인, 1995, pp. 91-98), 지리학(Kruskal과 Wish, 1978, pp. 7-9; Johnson과 Wichern, 2002, pp. 700-708), 교육학(du Toit 외 2인, 1984, pp. 133-144), 형상 및 이미지분석(Dryden과 Mardia, 1998, Chapter 12; Bronstein 외 2인, 2006, 1168-1172) 등 여러 분야에 다양하게 응용되고 있다.

1.3 다차원척도법의 이론적 배경

계량형과 비계량형을 포함하는 이원 다차원척도법과 삼원 다차원척도법의 본격적인 이론적 배경은 3장과 4장에서 설명하기로 하자. 이 절에서는 Kruskal과 Wish(1978, pp. 15-27), Everitt와 Dunn(1991, 5장), 그리고 허명회(1994, 8장)의 내용을 중심으로 다차원척도법의 이론적 배경을 간단히 소개하기로 한다.

p개의 변수(variables)로 n개의 개체(objects)에 대해 얻은 다변량자료행렬(multivariate data matrix)을 $X=(x_{ir})$, $i=1,...,n; r=1,...,p$라고 하자. X의 i번째 행벡터 $\boldsymbol{x}_i=(x_{i1}, ...,x_{ip})^t$는 i번째 개체를 나타내며 이들 n개의 행벡터 $\boldsymbol{x}_1,...,\boldsymbol{x}_n$은 p차원의 공간에서 좌표점으로 생각될 수 있다.

다차원척도법에서는 p차원의 공간에서 이들 좌표점 간의 거리(distance)가 매우 중요한 역할을 한다. 여기에서 공간이란 일반적으로 두 개체 간의 좌표점 $\boldsymbol{x}_i$와 $\boldsymbol{x}_j$ 사이의 거리가 피타고라스정리(Pythagorean formular)에 의한 유클리드거리

$$d_{ij} = \left[\sum_{r=1}^{p} (x_{ir} - x_{jr})^2\right]^{1/2} \tag{1.3}$$

로 정의되는 p차원의 유클리드공간을 말한다. 이 유클리드거리 d_{ij}를 1.1절의 식 (1.1)과 같이 비유사성으로 정의한다면 대각원소가 0이며 대칭인 비유사성행렬 D는 유클리드거리행렬이 된다. 따라서 다차원척도법이란 p차원의 유클리드공간에서 두 개체 간의 좌표점 $\boldsymbol{x}_i$와 $\boldsymbol{x}_j$ 사이의 거리 d_{ij}와 차원 축소된 $k(\leq p)$차원의 공간에서 두 개체 간의 거리 δ_{ij} 사이의 관계가 서로 일치되도록 하는 것이다. 보다 광범위하게 다차원척도법은 유클리드거리에 의해 정의되는 공간 외에 2장에서 설명되는 다양한 거리에 의한 다양한 공간에서 활용된다.

1.3.1 측정수준

다차원척도법에서 식 (1.3)에서 정의된 고차원의 거리 d_{ij}와 저차원의 δ_{ij}의 관계는 d_{ij}의 측정척도(measurement scale)에 따라 달라진다. Schiffman 외 2인(1981, p.15)과 Young(1987, pp.57-60)은 측정척도를 측정수준(measurement level)이라고 하여 다음과 같이 크게 나눈다.

- 명목수준(nominal level): 측정대상의 특성을 분류하기 위한 척도
  예) 성별, 동물의 성(genus), 사람의 머리카락 색깔, 음식의 맛

- 순서수준(ordinal level): 측정대상 간의 순서를 나타내는 척도
  예) 선호도, 부모의 수입 수준

- 수치수준(numerical level): 구간(interval)과 비율(ratio) 수준 모두를 의미하며 실제 수 자체가 수치적 의미를 지니고 있다.

수치수준에서 구간수준이란 등구간만 인정하는 경우로 예를 들면 화씨(F)온도에서 20도와 50도의 차이는 50도와 80도 차이와 같다. 그러나 80도가 20도보다 4배 덥다는 것을 의미하지는 않는다. 그리고 비율수준은 양(mass), 속도(velocity), 그리고 길이(cm)와 같은 측도와 관련이 있다. 시속 60km로 달리는 자동차는 시속 30km로 달리는 자동차보다 2배 빠르게 달린다고 볼 수 있다.

Everitt와 Dunn(1991, 5장)과 허명회(1994, 8장)는 이 측정수준에 근거하여 다음의 몇 가지 모형으로 d_{ij}와 δ_{ij}의 관계를 고려하였다.

1) 절대척도: $d_{ij} = \delta_{ij} + \epsilon_{ij}$,

2) 구간척도: $d_{ij} = a + b\delta_{ij} + \epsilon_{ij},\ b > 0$,

3) 비율척도: $d_{ij} = b\delta_{ij} + \epsilon_{ij},\ b > 0$,

4) 순서척도: $d_{ij} = f(\delta_{ij}) + \epsilon_{ij}$.

여기서 ϵ_{ij}는 측정 및 $k(\leq p)$차원 공간근사에 따른 왜곡오차이며 대개 오차항이라고 하고 $f(\cdot)$는 단조함수이다. 절대척도란 측정 거리값을 그대로 받아들이는 경우이다. 끝으로 3)의 비율척도모형은 구간척도모형 2)에서 $a = 0$인 경우에 해당한다.

이런 이유에서 Kruskal과 Wish(1978, pp. 22-23)는 d_{ij}와 δ_{ij}의 관계가 절대척도, 구간척도, 비율척도모형을 만족하도록 하는 다차원척도법을 계량형 다차원척도법이라고 하였고 순서척도모형의 경우는 비계량형 다차원척도법라고 하였다. 그리고 이들을 광범위하게는 이원 다차원척도법이라 하며 3장에서 이들을 위한 대표적인 알고리즘과 그 사례를 자세히 다루고 있다.

앞서 언급한 4가지 모형으로 정의한 d_{ij}와 δ_{ij}의 관계는 임의의 함수 $f(\cdot)$에 대해서 일반적으로

$$d_{ij} = f(\delta_{ij}) + \epsilon_{ij} \tag{1.4}$$

로 나타낼 수 있다. 예를 들면 식 (1.4)에서 함수 $f(\cdot)$를 선형인 $f(\delta_{ij}) = a + b\delta_{ij}$로 두면 구간척도모형이 된다.

1.3.2 스트레스

식 (1.4)에 있는 일반적 척도모형에서 d_{ij}와 $f(\delta_{ij})$의 관계를 최적화시킬 $f(\cdot)$를 찾기 위해서는 스트레스(stress)를 정의하고 이를 최소화시키는 수치적 알고리즘을 살펴보기로 하자. 이를 위해 먼저 Kruskal과 Wish(1978, pp. 24-25)의 잔차제곱합(residual sum of squares)

$$RSS = \sum_{i<j}\sum \epsilon_{ij}^2 = \sum_{i<j}\sum (d_{ij} - f(\delta_{ij}))^2 \tag{1.5}$$

을 생각하자. 척도변환에 이 잔차제곱합이 불변하도록 척도인자(scale factor)

$$SC = \sum_{i<j}\sum d_{ij}^2 \tag{1.6}$$

를 나누어 그 결과의 제곱근을 취하면 소위 스트레스(stress)

$$ST1 = (RSS/SC)^{1/2} = \left[\sum_{i<j}\sum (d_{ij} - f(\delta_{ij}))^2 \Big/ \sum_{i<j}\sum d_{ij}^2\right]^{1/2}$$

(1.7)

을 얻게 된다.

Kruskal과 Wish(1978, pp. 24-26)는 식 (1.7)의 스트레스를 〈objection function, criterion function, error function, evaluation function, merit function, goodness-of-fit function, badness-of-fit function〉으로 부른다. 일반적으로 식 (1.7)

의 $ST1$을 Kruskal의 〈스트레스 공식 1(stress formular 1)〉 또는 〈스트레스 1〉로 부르고 있다. 이런 의미에서 Kruskal의 〈스트레스 2〉는

$$ST2 = \left[\sum_{i<j}\sum (d_{ij} - f(\delta_{ij}))^2 \;/\; \sum_{i<j}\sum (d_{ij} - \bar{d}_{..})^2\right]^{1/2} \tag{1.8}$$

로 정의된다. 이는 식 (1.7)에서 사용된 식 (1.6)의 척도인자 SC 대신에 총 평균 $\bar{d}_{..} = \sum\sum d_{ij}/n^2$ 에 대해 수정된 척도인자 $\sum\sum (d_{ij} - \bar{d}_{..})^2$를 사용한 것이다.

스트레스는 이외에도 SAS/PROC ALSCAL에서 채택된 Young의 S-스트레스 공식 1과 2가 있다. 이는 Kruskal의 $ST1$과 $ST2$에 대응해서

$$SST1 = \left[\sum_{i<j}\sum (d_{ij}^2 - f(\delta_{ij})^2)^2 \;/\; \sum_{i<j}\sum (d_{ij}^2)^2\right]^{1/2} \tag{1.9}$$

$$SST2 = \left[\sum_{i<j}\sum (d_{ij}^2 - f(\delta_{ij})^2)^2 \;/\; \sum_{i<j}\sum (d_{ij}^2 - \bar{d}_{..}^2)^2\right]^{1/2} \tag{1.10}$$

로 정의된다(Young과 Lewyckyj, 1979, Chapter 4; Davison, 1992, pp. 87-89).

전체적으로 스트레스 공식을 요약하면 Kruskal의 $ST1$과 $ST2$ 그리고 Young의 $SST1$과 $SST2$의 차이점은 d_{ij}와 $f(\delta_{ij})$의 관계를 최적화시킬 $f(\cdot)$와 d_{ij}^2와 $f(\delta_{ij})^2$의 관계를 최적화시킬 $f(\cdot)$를 찾느냐에 달려 있다. 끝으로 척도인자를 나눌 때 평균에 대하여 식 (1.7)과 식 (1.9)와 같이 수정되지 않은 것을 공식 1(formula 1)이라고 하여 $ST1$과 $SST1$으로 식 (1.8)과 식 (1.10)과 같이 수정된 것은 공식 2(formular 2)이라고 하여 $ST2$과 $SST2$로 정의된다.

[알고리즘 1] 스트레스 최소화 알고리즘

1단계: p차원 유클리드공간에서 개체점 좌표 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$과 d_{ij}들에 대해 스트레스를 최소화하는 최적의 $f(\cdot)$를 찾는 것이 δ_{ij}의 최적변환을 찾는 것이다.

· 구간척도: $f(\delta_{ij}) = a + b\delta_{ij}$ 에서 최적의 a와 b를 구하는 것이다.

· 비율척도: $f(\delta_{ij}) = b\delta_{ij}$에서 최적의 b를 구하는 것이다.

· 순서척도: $f(\delta_{ij})$가 단조함수로 최적의 $f(\cdot)$를 구하는 것이다.

· 절대척도: $f(\delta_{ij}) = \delta_{ij}$이므로 $f(\cdot)$를 구할 필요가 없다.

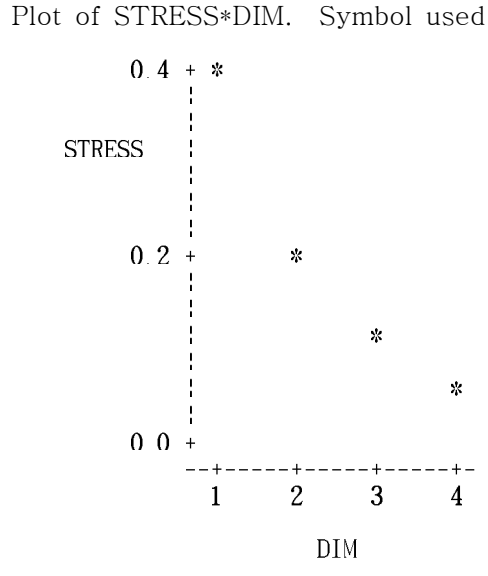
2단계: d_{ij}에 대한 측정모형에서 δ_{ij}의 최적변환에 조건화하여 스트레스를 최소화하는 k차원 좌표점을 구한다. 이때 최경사하강법 가우스-뉴튼법(Gauss-Newton method), Levenberg-Marquardt법과 같은 수치최적화 알고리즘이 적용된다.

따라서 이들 스트레스를 최소화하는 최적의 $f(\cdot)$를 찾는 것이 d_{ij}와 $f(\delta_{ij})$의 관계를 최적화한다. 일반적으로 구간과 비율척도에서는 회귀모형에서 가장 많이 쓰이는 최소제곱선형회귀(least squares linear regression) 알고리즘이 적용되고, 순서척도에는 최소제곱단조회귀(least squares monotonic regression) 알고리즘이 적용된다. 절대척도는 이 단계를 거치지 않는다. 이는 [알고리즘 1]의 두 단계로 구성된다(Kruskal과 Wish, 1978, pp. 25-27; Everitt와 Dunn, 1991, pp. 75-76; 허명회, 1994, 8장). 특히, [알고리즘 1]의 [2단계]의 가우스-뉴튼법과 Levenberg-Marquardt법은 비선형회귀(nonlinear regression) 문제를 해결하는 수치적 방법으로 자세한 수리적 설명과 비교는 Kennedy와 Gentle(1980, pp. 475-485)을 참고하기를 바란다.

1.3.3 차원의 수

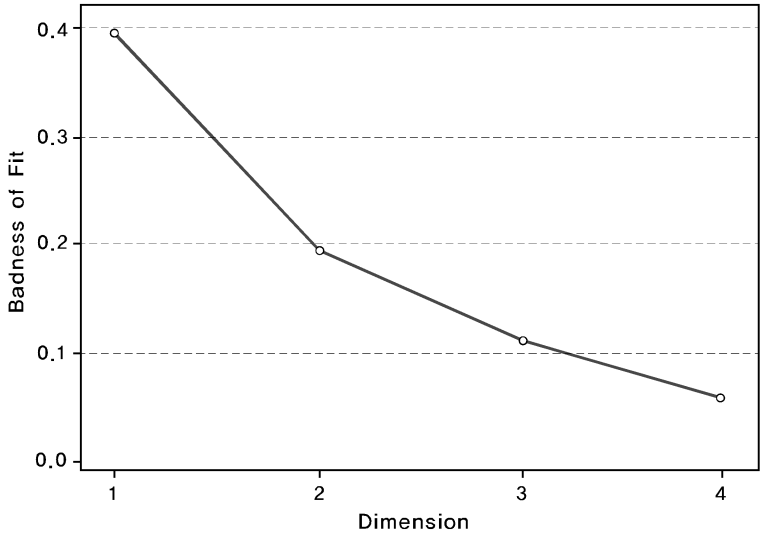
일반적으로 n개의 개체가 표현될 저차원의 차원수 $k(\leq p)$는 궁극적으로 다차원척도그림의 좌표축의 수를 정하는 문제와도 관련이 있다. Kruskal과 Wish(1978, Chapter 3)는 스트레스가 차원의 수가 증가하면 항상 감소하므로 이들의 그림을 통하여 스트레스가 제일 많이 꺾이는 곳에 대응하는 차원을 정하도록 제안하고 있다. 여기에서는 이를 스트레스-차원수 그림이라고 부르자. 더군다나 이 방법은 Catell(1966)에 인자분석(factor analysis)에서 인자의 개수를 정할 때 활용하는 〈scree plot〉과 닮았고 참고로 SAS/PROC FACTOR에서 〈scree〉 옵션으로 쉽게 제공되기도 한다. 그러나 6.07판의 SAS/PROC MDS에서는 스트레스-차원수 그림이 제공되는 옵션이 없어 최용석(1995)은 PROC MDS문에서 〈NOPHIST DIM=1 TO 4 OCRIT OUT= GRAPH〉을 사용하고 PROC PLOT 문을 사용하여 1차원부터 4차원까지의 스트레스와 차원수에 대한 〈그림 1.6〉과 같은 스트레스-차원수 그림을 제공 받을 수 있음을 보여 주었다. 이들 옵션에 대해선 [부록 1: SAS MDS 절차]에서

자세히 다루겠지만 1차원부터 4차원까지(DIM=1 to 4)의 마지막 반복에서의 부적합도 기준인 스트레스값(OCRIT)을 GRAPH라는 새로운 자료세트로 저장하되 스트레스 값을 찾는 반복과정은 생략(NOPHIST)하라는 것이다.



〈그림 1.6〉 6.07판의 스트레스-차원수 그림

이 책에서는 9.2판인 SAS Institute Inc.(2008, Chapter 53)에서는 그래픽 기능을 강화한 ODS GRAPHICS 문을 활용하고 PROC MDS 문에서 〈NOPHIST DIM=1 TO 4〉을 사용하면 〈그림 1.7〉과 같은 스트레스-차원수 그림을 제공받을 수 있다. 〈그림 1.6〉과 〈그림 1.7〉을 보면 2차원에서 다소 꺾임을 볼 수 있다. 〈그림 1.7〉은 실제로 1.4.3절의 [사례 3]에서 [프로그램 1.3]의 수행에서 얻을 수 있다.



〈그림 1.7〉 9.2판의 스트레스-차원수 그림

일반적으로 Chatfield와 Collins(1980, p. 208)는 2차원-4차원 정도에서 적절한 차원을 찾도록 권하고 있다. 대개 저자의 경험상 2차원(좌표축이 2개) 정도이면 차원축소를 목표로 하는 다변량기법(주성분분석, 인자분석, 행렬도, 대응분석 등)의 기하적해석이 원만하게 이루어진다고 생각된다.

이외에도 차원수 결정을 위한 적합도(goodness of fit)로 잘 알려진 크루스칼의 판별기준은 〈표 1.9〉와 같다(Kruskal, 1964; Mardia 외 2인, 1979, p. 414; Everitt와 Dunn, 1991, p. 77; Johnson과 Wichern, 2002, p. 701; Izenman, 2008, p. 501).

〈표 1.9〉 차원수 결정을 위한 크루스칼의 판별기준

스트레스	0.0	0.025	0.05	0.10	0.20
적합도	완벽 (perfect)	뛰어남 (excellent)	좋음 (good)	보통 (fair)	나쁨 (poor)

1.3.4 다차원척도그림의 해석

지금까지 다차원척도법의 이론적 배경을 간단해 살펴보았다. 일반적으로 p차원의 유클리드공간에 나타나는 개체 간의 거리 d_{ij}가 측정수준에 따라 차원축소된 $k(\leq p)$차원의 공

간에서 개체 간의 거리 δ_{ij}와의 관계가 일치되도록 [알고리즘 1]을 통하여 스트레스를 최소화하는 좌표점을 얻게 된다. 차원의 수는 부적합도 기준인 스트레스 값을 활용한 스트레스-차원수 그림과 크루스칼의 판별기준에 의해서 결정될 수 있으나 보편적으로 2차원이 선호된다. 1.1.4절의 [사례 4]에는 차원의 수는 3이지만 차원별로 짝을 지어 〈그림 1.4〉와 같이 15군데 신체 부위의 유사성에 대한 2차원의 다차원척도그림의 해석을 하였다.

일반적으로 2차원 다차원척도그림에 대한 해석은 [사례 1]-[사례 5]에서 보았듯이 크게 세 가지이다.

1) 개체 간의 그룹: 공통의 특성을 갖는 개체들은 가깝게 모여 있다.
2) 개체 간의 순서: 중요한 특성(좌표축)에 따라 개체들이 정렬된다.
3) 축에 대한 해석: 개체들의 그룹화와 순서화에 따라 좌표축을 해석하거나 새로운 축을 형성한다. 이 경우 축의 회전도 고려된다.

특히, 3)의 축에 대한 해석에서 새로운 축은 기존의 좌표축을 회전시켜 얻어진다고 볼 수 있다. 이는 인자분석에서 명확하지 못한 인자해석을 인자적재행렬의 회전을 통하여 새로운 축을 얻어 명확하게 하는 개념과 유사하다. 이러한 기하적인 설명은 개체 간에 정의한 유클리드거리에 근거한다고 볼 수 있다. 추가적으로 다양한 예제를 통한 상세한 다차원척도그림에 대한 해석은 Davison(1992, pp. 70-71과 pp. 93-94)을 참고하기를 바란다.

1.4 다차원척도법을 위한 SAS/PROC MDS의 예

이 절에서는 1.1절에서 소개한 5가지 사례 중 [사례 1]-[사례 3]과 관련된 SAS/PROC MDS 프로그램을 소개하고 나머지 [사례 4]와 [사례 5]의 프로그램은 각각 4장과 5장에서 소개하려 한다.

1.4.1 [사례 1]을 위한 프로그램 및 결과

[프로그램 1.1]은 1.1.1절의 〈표 1.1〉의 도시 간 철도거리에 대한 다차원척도법을 위한 SAS/PROC MDS 프로그램이다. [프로그램 1.1]은 크게 두 과정(procedure)으로 나뉘어져 있다. 먼저 1)은 〈표 1.1〉의 자료를 SAS 자료세트인 〈railroad〉를 만드는 과정이다. 2)는

다차원척도법을 위한 PROC MDS 문이다. 자료의 측정수준이 절대척도이므로 옵션 〈level=absolute〉를 사용하였다. 따라서 이 옵션은 계량형 다차원척도법을 암시한다. 측정수준에 대해선 1.3.1절에서 충분히 언급하였다. 그리고 〈pconfig〉는 [결과 1.1]의 ①과 ②만을 제공하라는 출력에 관련된 옵션이다.

[프로그램 1.1] 도시 간 철도거리의 계량형 다차원척도법

```
1) data railroad;
        input (seoul pusan kwangju taegu
                incheon kangnung cheongju jeonju) (6.1) @60;
  cards;
     0.0
   444.5     0.0
   353.8   364.7     0.0
   323.9   120.6   350.0     0.0
    30.9   475.4   384.7   359.9     0.0
   348.5   505.0   523.0   388.8   379.4     0.0
   141.2   326.1   235.4   211.4   172.1   288.6     0.0
   279.9   391.0   124.6   276.4   310.8   449.4   136.3   0.0
     ;
2) ods graphics on;
   proc mds data=railroad level=absolute pconfig;
```

결과 ①은 1.3.2절에서 설명한 [알고리즘 1]인 스트레스 최소화 알고리즘을 위한 가우스-뉴튼법과 Levenberg-Marquardt법을 적용한 수치적 계산과정을 나타낸다. 여기서 마지막 반복(iteration) 2의 〈Badness-of-Fit Criterion〉 0.0800은 스트레스 값이다. 이는 약 0.08로 결과 ③의 2차원 계량형 다차원척도그림이 개체 간의 관계를 설명하는 정도가 크루스칼의 판별 기준인 〈표 1.9〉에 따르면 〈0.05 좋음(good)〉보다 낮고 〈0.10 보통(fair)〉보다는 높은 정도를 나타낸다. 차원수는 디폴트(DIM=2)로 2이다. 그리고 결과 ②의 〈Configuration〉은 결과 ③의 형상공간인 2차원 다차원척도그림을 위한 좌표이며 이를 형상좌표(configuration coordinate)라고 한다.

끝으로 최용석(1995)의 SAS 6.07판에서 결과 ③의 다차원척도그림을 얻기 위해서는 다소 복잡한 PROC PLOT문을 활용하였지만, 9.2판부터는 [프로그램 1.1]의 2)에서 PROC MDS 문에 앞서 〈ods graphics on;〉을 삽입하면 자동으로 제공된다.

기본적으로 [부록 1.2]의 ODS GRAPHICS 절차에 관련된 내용에 대한 약간의 지식이 필요하지만 우선은 PROC MDS 절차에 앞서 [부록 1.2] ODS GRAPHICS의 [프로그램 2]를 제일 먼저 실행하면 된다. 이는 저자가 다차원척도그림에서 수평과 수직축의 원점을 지나는 참고 선을 그리고 그림 모양을 정사각형으로 제공하도록 추가적으로 수정한 것이다.

[결과 1.1] [프로그램 1.1]의 수행결과

①

Multidimensional Scaling: Data=WORK.RAILROAD.DATA
Shape=TRIANGLE Condition=MATRIX Level=ABSOLUTE
Coef=IDENTITY Dimension=2 Formula=1 Fit=1
Gconverge=0.01 Maxiter=100 Over=1 Ridge=0.0001

Iteration	Type	Badness-of-Fit Criterion	Change in Criterion	Convergence Measure
0	Initial	0.0944	-	0.5122
1	Lev-Mar	0.0802	0.0142	0.0610
2	Gau-New	0.0800	0.000167	0.008376

②

Convergence criterion is satisfied.

Configuration		
	Dim1	Dim2
Seoul	-145.21	-100.68
Pusan	207.98	205.76
Kwangju	196.72	-156.14
taegu	94.08	146.41
incheon	-169.99	-128.31
kangnung	-260.46	192.24
cheongju	-22.18	-29.37
jeonju	99.05	-129.90

③ 〈그림 1.1〉 도시 간 철도거리의 계량형 다차원척도그림

1.4.2 [사례 2]를 위한 프로그램 및 결과

[프로그램 1.2]는 (a)와 (b)로 구성되어 있고 그 수행 결과도 이와 대응하여 [결과 1.2]에 (a)와 (b)로 정리되어 있다.

1.1.2절의 〈표 1.2〉 우리나라 14개 경제관련기관의 10가지 경제 전망으로부터 (a)는 〈표 1.3〉 원자료의 유클리드거리에 의한 비유사성자료에 대한 다차원척도 프로그램이다. (b)는 〈표 1.4〉의 이진수 자료로부터 〈표 1.5〉의 비유사성행렬자료에 대한 다차원척도법을 위한 프로그램이다. 특히 〈표 1.5〉의 자료는 이진수 자료인 〈표 1.4〉에서 두 기관별 제곱유클리드거리에다가 경제 전망의 항목수인 10을 나눈 것과 동일하며, 이는 2장의 유사성 측도인 단순매칭계수(simple matching coefficient)와 연관되어 있다.

[프로그램 1.2]는 [사례 1]의 [프로그램 1.1]과 그 문법적 형식이 동일하다. 단지 두드러진 차이점은 PROC MDS 문에서 옵션 〈level= ordinal〉(디폴트)와 〈pfinal〉이다. 이는 비계량형 다차원척도법을 적용하기 위한 옵션이다.

[결과 1.2]-①은 비계량형 다차원척도법을 위한 수치적 계산과정이다. 이는 1.3.2절에서 설명한 $ST1$(Kruskal의 스트레스 1)의 최소화 알고리즘을 위한 것으로 [결과 1.1]-①에 비해 [알고리즘 1]의 스트레스 최소화 알고리즘의 [1단계]와 [2단계]를 다 거치고 있음을 보여 주고 있다.

옵션 〈pfinal〉은 옵션 〈pconfig〉에 의해서 제공되는 결과 ②뿐만 아니라 결과 ③도 제공하는 출력에 관련된 옵션이다. 결과 ③은 결과 ①-(a)와 (b)의 마지막 반복의 〈Badness-of-Fit Criterion〉값 0.1496과 0.1639만을 제공하고 있다. 이는 각각 약 0.15와 0.16에 해당하여 두 경우가 대동소이하고 크루스칼의 판별 기준 〈표 1.9〉로부터 〈0.10 보통(fair)〉보다는 낮고 〈0.20 나쁨(poor)〉보다는 높다는 것을 알 수 있다. 결과 ④는 〈그림 1.2〉 경제관련기관의 경제 전망의 비계량형 다차원척도그림으로 그 해석도와 함께 이미 1.1.2절에서 먼저 인용하여 살펴보았다. 이 외의 [프로그램 1.2]와 [결과 1.2]에 대한 나머지 설명은 1.4.1절의 [사례 1]에서와 같다.

[프로그램 1.2] 경제관련기관의 경제 전망에 관한 비계량형 다차원척도법

(a) 원자료의 유클리드거리

```
data economy_euclid;
        input (hankook daewoo1 dongseo1 kyungrun sangkong
             choongso hyundai samsung sunkyung daewoo2
             shinhan dongseo2 korea chungang) (6.3) @84;
cards;
 0.000
 4.166  0.000
 2.293  3.672  0.000
 3.269  6.925  4.110  0.000
 2.390  5.476  3.496  4.157  0.000
 4.410  5.723  4.635  5.837  4.316  0.000
 3.573  4.432  4.084  4.812  5.177  4.231  0.000
 1.846  4.197  2.839  4.322  3.182  5.220  4.233  0.000
 3.700  5.189  2.854  5.207  4.631  5.231  5.088  3.880  0.000
 3.131  6.345  3.800  2.789  3.820  4.796  3.928  4.468  4.159  0.000
 3.118  4.344  2.363  4.653  4.265  4.260  3.592  3.920  2.990  3.549  0.000
 1.784  4.443  1.515  3.291  3.061  4.263  3.820  2.469  3.401  3.278  2.648  0.000
 5.362  5.571  5.684  6.926  5.856  5.563  4.699  6.336  5.696  5.349  4.994  6.088  0.000
 4.343  6.693  5.425  4.474  5.124  4.407  3.336  5.114  6.064  4.192  4.934  4.568  5.176  0.000
 ;
ods graphics on;
proc mds data=economy_euclid pfinal;
```

(b) 이진수 자료의 제곱유클리드거리

```
data economy;
     input (hankook daewoo1 dongseo1 kyungrun sangkong
            choongso hyundai samsung sunkyung daewoo2
            shinhan dongseo2 korea chungang) (3.1) @45;
cards;
.0
.4  .0
.4  .2  .0
.4  .8  .8  .0
.5  .5  .5  .5  .0
.5  .3  .5  .7  .4  .0
.8  .4  .6  .6  .5  .3  .0
.3  .3  .5  .5  .4  .6  .5  .0
.6  .2  .2  .8  .5  .5  .6  .5  .0
.4  .8  .8  .0  .5  .7  .6  .5  .8  .0
.6  .2  .4  .6  .7  .5  .4  .5  .2  .6  .0
.3  .3  .1  .7  .4  .4  .7  .6  .3  .7  .5  .0
.8  .6  .8  .4  .5  .5  .2  .5  .6  .4  .4  .9  .0
.7  .5  .7  .5  .4  .2  .1  .6  .7  .5  .5  .6  .3  .0
 ;
     ods graphics on;
proc mds data=economy pfinal;
```

[결과 1.2] [프로그램 1.2]의 수행결과

(a) 원자료의 유클리드거리

Shape=TRIANGLE Condition=MATRIX Level=ORDINAL
Coef=IDENTITY Dimension=2 Formula=1 Fit=1
Mconverge=0.01 Gconverge=0.01 Maxiter=100 Over=2
Ridge=0.0001

①

Iteration	Type	Badness-of-Fit Criterion	Change in Criterion	Convergence Measure	
				Monotone	Gradient
0	Initial	0.2776	.	.	.
1	Monotone	0.2049	0.0727	0.1760	0.5150
2	Gau-New	0.1708	0.0341	.	.
3	Monotone	0.1656	0.005190	0.0336	0.2572
4	Gau-New	0.1598	0.005814	.	.
5	Monotone	0.1532	0.006609	0.0436	0.1787
6	Gau-New	0.1525	0.000733	.	.
7	Monotone	0.1513	0.001152	0.0174	0.1406
8	Gau-New	0.1508	0.000518	.	.
9	Monotone	0.1503	0.000486	0.0117	0.0958
10	Gau-New	0.1501	0.000215	.	.
11	Monotone	0.1500	0.000130	0.006270	0.0726
12	Gau-New	0.1496	0.000379	.	0.006127

(b) 이진수 자료의 제곱유클리드거리

Shape=TRIANGLE Condition=MATRIX Level=ORDINAL
Coef=IDENTITY Dimension=2 Formula=1 Fit=1
Mconverge=0.01 Gconverge=0.01 Maxiter=100 Over=2
Ridge=0.0001

Iteration	Type	Badness-of-Fit Criterion	Change in Criterion	Convergence Measure	
				Monotone	Gradient
0	Initial	0.2275	.	.	.
1	Monotone	0.2144	0.0131	0.0813	0.5365
2	Gau-New	0.1764	0.0380	.	.
3	Monotone	0.1759	0.000421	0.0331	0.2770
4	Gau-New	0.1702	0.005737	.	.
5	Monotone	0.1671	0.003085	0.0359	0.1912
6	Gau-New	0.1668	0.000327	.	.
7	Monotone	0.1668	0.0000296	0.008016	0.1847
8	Gau-New	0.1639	0.002879	.	0.007368

②

Configuration		
	Dim1	Dim2
hankook	0.40	0.06
daewoo1	-1.11	1.75
dongseo1	0.21	0.62
kyungrun	1.59	-0.78
sangkong	1.22	0.05
choongso	-0.88	-1.33
hyundai	-0.84	-0.32
samsung	0.81	0.74
sunkyung	0.40	1.45
daewoo2	0.67	-0.77
shinhan	-0.26	0.47
dongseo2	0.42	0.23
korea	-2.54	-0.28
chungang	-0.09	-1.89

Configuration		
	Dim1	Dim2
hankook	0.33	1.55
daewoo1	0.96	-0.25
dongseo1	1.49	0.38
kyungrun	-1.60	0.92
sangkong	-0.48	0.34
choongso	0.09	-1.04
hyundai	-0.70	-1.28
samsung	-0.06	0.99
sunkyung	1.49	-0.32
daewoo2	-1.60	0.92
shinhan	0.93	-0.93
dongseo2	1.36	0.62
korea	-1.41	-0.83
chungang	-0.81	-1.07

③

MATRIX	Number of Nonmissing Data	Weight	Badness-of-Fit Criterion	Distance Correlation	Uncorrected Distance Correlation
1	91	1.00	0.15	0.94	0.99

MATRIX	Number of Nonmissing Data	Weight	Badness-of-Fit Criterion	Distance Correlation	Uncorrected Distance Correlation
1	91	1.00	0.16	0.91	0.99

④ 〈그림 1.2〉 경제관련기관의 경제 전망의 비계량형 다차원척도그림

1.4.3 [사례 3]을 위한 프로그램 및 결과

[프로그램 1.3]은 1.1.3절의 12개 국가의 유사성행렬자료 〈표 1.6〉을 이용하여 SAS자료세트인 〈nation〉을 만드는 과정 1)과 다차원척도법을 위한 PROC MDS 문의 과정 2)로 구성되어 있다. 프로그램의 수행결과는 [결과 1.3]에 정리되어 있다.

기본적으로 SAS/PROC MDS에서는 비유사성행렬자료에 대한 다차원척도법을 제공한다. 따라서 유사성행렬자료를 비유사성행렬자료로 바꾸어 주는 과정 내지는 이를 위한 옵션이 필요하다. 다시 말해서 먼저 1.1.3절에서 살펴본 바와 같이 개체 자신의 유사성이 9점(매우 유사하다)을 대각원소로 하는 유사성행렬자료 〈표 1.6〉을 입력자료로 하여 전형적인 PROC MDS 문을 사용한다. 그러면 PROC MDS는 자동으로 대각원소값이 비대각원소값보다 크므로 대각원소값에서 비대각원소값을 빼 비유사성행렬로 만든다.

[프로그램 1.3]의 2)에서 측정수준에 관한 옵션으로 〈level=ordinal〉(디폴트)로 순서척도를 지정하여 비계량형 다차원척도법을 제공하고 있다. 이는 〈표 1.4〉의 유사성 행렬자료에서 원소들의 크기의 순서에 의미를 갖기 때문이다. 옵션 〈nophist〉는 이미 살펴본 [사례 1]과 [사례 2]의 [결과 1.1]과 [결과 1.2]의 ①에 나타난 스트레스 최소화 알고리즘을 위한 수치적 반복과정을 제공하지 않는다. 대신에 제일 마지막 〈Badness-of-Fit Criterion〉의

스트레스 값 0.20만을 [결과 1.3]에서 보여 주고 있다. [프로그램 1.3]의 과정 2)의 〈ods graphics on;〉으로 인하여 자동으로 [결과 1.3]의 〈그림 1.3〉을 제공한다. 그리고 옵션 〈dim=1 to 4〉에 의해서 1.3.3절에서 언급한 형상공간의 차원의 수를 결정할 스트레스-차원수 그림 〈그림 1.7〉을 제공한다.

[프로그램 1.3] 12개 국가의 유사성행렬자료에 대한 비계량형 다차원척도법

```
1) data nation;
        input (brazil congo cuba egypt france india
                israel japan china ussr usa yugo) (5.2) @60;
   cards;

   9.00
   4.83  9.00
   5.28  4.26  9.00
   3.44  5.00  5.17  9.00
   4.72  4.00  4.11  4.78  9.00
   4.50  4.83  4.00  5.83  3.44  9.00
   3.83  3.33  3.61  4.67  4.00  4.11  9.00
   3.50  3.39  2.94  3.84  4.11  4.50  4.83  9.00
   2.39  4.00  5.50  4.39  3.67  4.11  3.00  4.17  9.00
   3.06  3.39  5.44  4.39  5.06  4.50  4.17  4.61  5.72  9.00
   5.39  2.39  3.17  3.33  5.94  4.28  5.94  6.06  2.56  5.00  9.00
   3.17  3.50  5.11  4.28  4.72  4.00  4.44  4.28  5.06  6.67  3.56  9.00
   ;
2) ods graphics on;
   proc mds data=nation pfinal nophist dim=1 to 4;
```

[결과 1.3] [프로그램 1.3]의 수행결과

Multidimensional Scaling: Data=WORK.NATION.DATA
Shape=TRIANGLE Condition=MATRIX Level=ORDINAL
Coef=IDENTITY Dimension=2 Formula=1 Fit=1

Configuration		
	Dim1	Dim2
brazil	-0.34	1.78
congo	1.41	1.28
cuba	1.43	-0.27
egypt	0.79	0.31
france	-0.61	0.31
india	0.40	0.61
israel	-1.45	0.35
japan	-1.30	-0.78
china	1.26	-1.18
ussr	0.01	-1.23
usa	-1.66	0.03
yugo	0.07	-1.20

MATRIX	Number of Nonmissing Data	Weight	Badness-of-Fit Criterion	Distance Correlation	Uncorrected Distance Correlation
1	66	1.00	0.20	0.85	0.98

〈그림 1.3〉 12개 국가의 유사성에 대한 비계량 다차원척도그림
〈그림 1.7〉 9.2판의 스트레스-차원수 그림

이번에는 [프로그램 1.3]의 과정 1)에서 대각원소 9 대신에 · (결측값)이나 0.00(zero)을 입력하고 과정 2)의 PROC MDS 문에 〈similar=9〉 옵션을 추가하면 동일한 [결과 1.3]을 제공받을 수 있다. 여기서 〈similar=9〉는 입력자료를 비유사성행렬이 아닌 유사성행렬로 처리하라는 옵션이다. 그리고 지정된 9는 이 값과 자료행렬의 원소들 중 최댓값으로부터 각 원소를 빼 비유사성행렬을 제공하라는 수이다. 이와 관련하여 [부록 1.1.2]에서 유사성행렬자료를 입력자료로 사용할 경우에 대하여 자세히 설명하고 있다.

연습문제

1.1 다음은 우리나라 경부, 호남, 영동, 남해, 88올림픽고속도로 중심의 8개 도시 간 고속도로 거리(km)에 대한 자료이다(허명회, 1992).

서울	.							
대전	152	.						
전주	225	85	.					
광주	317	177	92	.				
대구	305	153	238	208	.			
부산	428	276	361	258	123	.		
원주	119	211	284	376	364	598	.	
강릉	230	322	395	487	475	709	111	.

1) 자료가 유사성행렬인지 또는 비유사성행렬인지를 말하라.
2) 대각원소값은 얼마인가?
3) 측정수준과 스트레스값을 구하라.
4) 스트레스-차원수 그림을 얻고 적절한 차원을 말하라.
5) 다차원척도그림을 해석하고 새로운 축에 의미를 부여하라.

1.2 다음은 미국 주요 10개 도시들 간의 비행거리(miles)를 나타낸다(Kruskal과 Wish, 1978, pp.7-9).

1) 비행거리를 바탕으로 가까운 도시를 군집할 때 동부(New york, Washington D.C., Atlanta, Chicago, Miami, Houston)와 서부(Los Angeles, San Francisco, Seattle, Denver)로 나누어지는지를 확인하여 보라.
2) 다차원척도법을 통하여 군집화를 시도하고 축을 해석하라.

3) 계량형과 비계량형 다차원척도법의 결과를 비교하라.

4) 스트레스값은 얼마이며 적합의 정도를 판단하라.

ATLANTA	0									
CHICAGO	587	0								
DENVER	1212	920	0							
HOUSTON	701	940	879	0						
LOS ANGELES	1936	1745	831	1374	1374					
MIAMI	604	1188	1726	968	2339	0				
NEW YORK	748	713	1631	1420	2451	1092	0			
SAN FRANCISCO	2139	1858	949	1645	347	2594	2571	0		
SEATTLE	2182	1737	1021	1891	959	2734	2408	678	0	
WASHINGTON D.C.	543	597	1494	1220	2300	923	205	2442	2329	0

1.3 다음의 자료는 4개의 기업(아남, 대우전자, 금성사, 삼성전자)에 대한 자료이다(최병진, 1989). 이 자료는 239명의 사람들이 각각 기업을 둘씩 짝지은 6가지 경우에 대해 1점(아주 비슷하다) ~ 7점(매우 다르다)의 7점 척도로 응답한 후 그 평균으로 만들어졌다.

아남	.			
대우전자	4.2585	.		
금성사	4.3178	4.1350	.	
삼성전자	4.4430	4.0000	3.6410	.

1) 자료가 유사성행렬인지 또는 비유사성행렬인지를 말하라.

2) 대각원소값은 얼마인가?

3) 스트레스-차원수 그림을 얻고 적절한 차원을 말하라.

4) 측정수준을 달리하여 다차원척도그림을 서로 비교하여 보라.

5) 다차원척도그림을 해석하고 개체 간의 그룹화와 그 특성을 논하라.

1.4 다음은 켈로그회사에 의해서 제조된 총 23종류의 시리얼에 대한 것으로 10가지 변수 $X_1 \sim X_{10}$으로 측정되었고 이들을 다시 표준화한 자료이다. 이는 [부록 2: 자료 5(파일이름: kellogg.txt)]에서 제공되며 변수들은 다음과 같다.

X_1: 칼로리, X_2: 단백질(g), X_3: 지방(g), X_4: 나트륨(mg), X_5: 다이어트 식이섬유(g), X_6: 복합탄수화물(g), X_7: 당분(g), X_8: 칼륨(mg), X_9: 비타민과 무기물(하루 권장량(%) : 0, 25, 100), X_{10}: 유형(온 또는 냉)

1) 주관적 기준에 따라 이진수 자료를 만들어 보라.
2) 두 개체 간에 일치하지 않는 항목의 수를 이용하여 비유사성행렬자료를 제공하라.
3) 비유사성행렬자료에 대한 다차원척도법을 실시하고 시리얼의 유형을 군집화하라.
4) 주성분분석을 통한 주성분점수의 산점도와 다차원척도그림을 비교하여 시리얼의 유형에 차이가 있는지를 살펴보라.

시리얼	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
AllB	0.1818	0.6	0.3333	0.8125	0.6429	0	0.3333	1	0.9677	0
AllF	0	0.6	0	0.4375	1	0.0667	0	1	1	0
AppJ	0.5455	0.2	0	0.3906	0.0714	0.2667	0.9333	0.5	0.0323	0
CorF	0.4545	0.2	0	0.9063	0.0714	0.9333	0.1333	0	0.0484	0
CorP	0.5455	0	0	0.2813	0.0714	0.4	0.8	0.5	0	0
Crac	0.5455	0.4	1	0.4375	0.2857	0.2	0.4667	1	0.4516	0
Cris	0.5455	0.2	0	0.6875	0.0714	0.9333	0.2	1	0.0323	0
Froo	0.5455	0.2	0.3333	0.3906	0.0714	0.2667	0.8667	0.5	0.0323	0
FroF	0.5455	0	0	0.625	0.0714	0.4667	0.7333	0	0.0161	0
FrMW	0.4545	0.4	0	0	0.2143	0.4667	0.4667	0.5	0.2581	0
FruB	0.6364	0.4	0	0.75	0.3571	0.4667	0.8	1	0.5484	0
JRCN	0.5455	0.2	0.3333	0.5313	0.0714	0.6667	0.4	1	0.129	1
JRFN	0.8182	0.4	0.3333	0.5313	0.1429	0.8667	0.6	1	0.2419	1
MuCB	1	0.4	0.6667	0.4688	0.2143	0.6667	0.8667	1	0.4516	0
Nut&	0.6364	0.2	0.3333	0.5938	0	0.5333	0.6	0.5	0.0645	0
NGAR	0.8182	0.4	0.6667	0.6875	0.2143	0.9333	0.4667	1	0.3548	0
NutW	0.3636	0.4	0	0.5313	0.2143	0.7333	0.1333	1	0.2258	0
Prod	0.4545	0.4	0	1	0.0714	0.8667	0.2	1	0.0806	1
RaBr	0.6364	0.4	0.3333	0.6563	0.3571	0.4667	0.8	0.5	0.7097	0
Rais	0.3636	0.2	0	0	0.1429	0.5333	0.4	1	0.2903	0
RiKr	0.5455	0.2	0	0.9063	0	1	0.2	0	0.0484	0
Smac	0.5455	0.2	0.3333	0.2188	0.0714	0.1333	1	0.5	0.0645	0
Spec	0.5455	1	0	0.7188	0.0714	0.6	0.2	0	0.1129	0

1.5 다음은 사람의 눈으로 색을 인지하는 실험의 결과이다. 색채의 파장 길이($434\mu m \sim 674\mu m$)에 따른 14가지 색을 각각 두 색씩 짝을 지은 91가지 경우에 대하여 31명이 0점(전혀 유사하지 않다) ~ 4점(동일하다)의 5점척도로 평가하여 그 평균 점수로 14가지 색의 행렬자료를 만들었다. 이는 [부록 2: 자료 6(파일이름: color.txt)]에서 제공되며 참고로 파장 길이에 따라 434=남색, 445=청색, 472=청록색, 504=녹색, 555=황녹색, 600=노랑색, 628=귤색, 651=주황색, 674=적색 등 몇 가지 색을 살펴볼 수 있다(Izenman, 2008, pp. 468-469).

1) 행렬자료가 유사성행렬인지 또는 비유사성행렬인지를 말하라.

2) 비계량형 다차원척도법에서 다차원척도그림은 색채환(color circle)을 제공함을 보여라.

3) 비계량형과 계량형 다차원척도법의 결과를 비교하라.

434	1													
445	0.86	1												
465	0.42	0.5	1											
472	0.42	0.44	0.81	1										
490	0.18	0.22	0.47	0.54	1									
504	0.06	0.09	0.17	0.25	0.61	1								
537	0.07	0.07	0.1	0.1	0.31	0.62	1							
555	0.04	0.07	0.08	0.09	0.26	0.45	0.73	1						
584	0.02	0.02	0.02	0.02	0.07	0.14	0.22	0.33	1					
600	0.07	0.04	0.01	0.01	0.02	0.08	0.14	0.19	0.58	1				
610	0.09	0.07	0.02	0	0.02	0.02	0.05	0.04	0.37	0.74	1			
628	0.12	0.11	0.01	0.01	0.01	0.02	0.02	0.03	0.27	0.5	0.76	1		
651	0.13	0.13	0.05	0.02	0.02	0.02	0.02	0.02	0.2	0.41	0.62	0.85	1	
674	0.16	0.14	0.03	0.04	0	0.01	0	0.02	0.23	0.28	0.55	0.68	0.76	1

2장
비유사성과 유사성

다차원척도법에서 이용되는 최종 자료의 형태는 식 (1.1)에서 정의한 크기가 $n \times n$인 비유사성행렬 $D = (d_{ij})$, $i, j = 1, \ldots, n$이거나 1.1.3절의 [사례 3]에서 정의한 유사성 c_{ij}에 의한 유사성행렬 $C = (c_{ij})$, $i, j = 1, \ldots, n$이다. 비유사성과 유사성은 경우에 따라 상호 변환(transformation)이 가능하다. 일반적으로 이 행렬 D와 C는 직접 관찰되기도 하지만 대개 1.3절에서 주어진 p개의 변수로 n개의 개체에 대해 얻은 다변량자료행렬 $X = (x_{ir}), i = 1, \cdots, n; r = 1, \cdots, p$로부터 구해진다. 이 장에서는 개체 간과 변수 간의 비유사성과 유사성에 대해 각각 정의하고 이를 측정할 측도에 대한 소개 및 적용사례를 보이고자 한다. 더불어 그룹 간의 비유사성계수와 양적변수, 질적변수가 혼합되어 있는 경우의 유사성계수를 추가적으로 설명하고 있다. 끝으로 원자료나 이진수자료로부터 비유사성과 유사성측도를 계산하고 이를 통하여 제공되는 비유사성과 유사성행렬자료의 다차원척도법을 통합적으로 적용하게 하는 프로그램을 소개하려 한다.

2.1 비유사성

비유사성(dissimilarity)은 개체 간의 근접성(proximity 또는 closeness)과 연관성(association)을 평가하는 측도이다. 이는 개체가 변수라면 상관계수의 절댓값에 의해서 측정되며, 일반적으로 유클리드거리, 네트워크거리 등 다양한 기하적거리를 나타낸다. 비유사성은 다차원척도법뿐만 아니라 군집분석에서도 활용되며, 이를 정의하고 요약하자.

2.1.1 비유사성 측도

다변량자료행렬 X의 i번째 행벡터 $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ip})^t$는 i번째 개체를 나타내며 이들 n개의 행벡터 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$은 p차원의 공간에서 좌표점이다. 따라서 다차원척도법에서는 p차원의 공간에서 이들 좌표점 간의 거리가 매우 중요한 역할을 하며 이를 비유사성의 측도로 사용한다.

두 개체 $\boldsymbol{x}_i$와 $\boldsymbol{x}_j$ 사이의 거리들 중에서 가장 잘 알려진 것이 식 (1.3)에서 정의하였지만

$$\textbf{유클리드거리}:\ d_{ij} = \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)^t(\boldsymbol{x}_i - \boldsymbol{x}_j)\right]^{1/2} = \left[\sum_{r=1}^{p}(x_{ir} - x_{jr})^2\right]^{1/2}$$

이다. 여기서 나머지 비유사성을 나타내는 거리들에 대해 정의하기에 앞서 이들이 기본적으로 만족해야 할 세 가지 조건은 다음과 같다.

1) $d_{ij} \geq 0,\ i \neq j$.
2) $d_{ii} = 0$.
3) $d_{ij} = d_{ji}$. (대칭성)

추가적으로 다음의 4a)를 만족하면 계량형(metric)이고, 4b)를 만족하면 초계량형(ultrametric)이라고 한다(Izenman, 2008, pp. 412-413).

4a) $d_{ik} + d_{jk} \geq d_{ij}$. (삼각 부등식)
4b) $\max\{d_{ik}, d_{jk}\} \geq d_{ij}$.

그러나 조건 4a)와 4b)는 결코 필요조건은 아니며 대부분의 거리들이 이 조건을 만족하지 않기 때문이다. 일반적으로 유클리드거리는 모든 조건을 만족한다고 알려져 있다. 다음으로는

$$\text{민코브스키(Minkowski)거리: } d_{ij} = \left[\sum_{r=1}^{p} | x_{ir} - x_{jr} |^k\right]^{1/k}, \ k \geq 1 \qquad (2.1)$$

가 있다. 식 (2.1)에서 $k=2$인 경우는 유클리드거리며 $k=1$인 경우는

$$\text{시티-블럭(city-block): } d_{ij} = \sum_{r=1}^{p} | x_{ir} - x_{jr} |$$

로 맨하턴(Manhattan)거리라고도 한다. 이들 세 가지 거리들은 변수들이 그들의 측정단위가 다르거나, 서로 상관관계가 있거나, 특정한 변수의 분산이 나머지 변수들에 비해 매우 큰 경우에 크게 달라진다(Chatfield와 Collins, 1980, pp. 192-193). 이 점을 개선하여 변수들이 중심화 $\Sigma_{i=1}^{n} x_{ir} = 0$, $r = 1,...,p$를 만족한다는 가정 아래 표준화를 통한 표준화거리로

$$\text{피어슨(Pearson)거리: } d_{ij} = \left[\sum_{r=1}^{p} (x_{ir} - x_{jr})^2 / s_r^2\right]^{1/2}$$

혹은 χ^2거리가 있다. 여기서 $s_r^2 = \Sigma_{i=1}^{n} (x_{ir} - \overline{x}_r)^2/(n-1)$은 r번째 변수 $\boldsymbol{x}_r = (x_{1r},...,x_{nr})^t$의 분산이다. 다음으로 피어슨거리가 변수별 표준화를 고려한 것이라면 이뿐만 아니라 변수들 간의 상관관계도 고려하고 있는

$$\text{마할라노비스(Mahalanobis)거리: } d_{ij} = \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)^t S^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)\right]^{1/2}$$

가 있다. 여기서 S는 표본분산-공분산 행렬이다. 끝으로 유클리드거리와 피어슨거리의 일반화된

$$\text{가중(weighted)유클리드거리: } d_{ij} = \left[\sum_{r=1}^{p} w_r \, (x_{ir} - x_{jr})^2 \right]^{1/2} \qquad (2.2)$$

를 생각할 수 있다. 식 (2.2)에서 $w_r = 1$이면 유클리드거리이고 $w_r = 1/s_r^2$이면 피어슨거리가 됨을 쉽게 알 수 있다.

Mardia 외 2인(1979, p. 381)은 그룹 간의 비유사성계수로 칼 피어슨과 마할라노비스 계수를 요약하고 있다. 만약에 임의의 두 그룹 h와 k에 대하여 각 그룹의 크기를 n_g, $g = h, k$라 하고 $\bar{x}_{gr}$과 s_{gr}^2을 g번째 그룹의 평균과 분산이라고 하면 그룹 간의 비유사성계수인 칼 피어슨거리

$$d_{hk} = \left[\frac{1}{p} \sum_{r=1}^{p} \frac{(\bar{x}_{hr} - \bar{x}_{kr})^2}{(s_{hr}^2/n_h) + (s_{kr}^2/n_k)} \right]^{1/2}$$

를 정의하며 마할라노비스거리로 다음을 정의하고 있다.

$$d_{hk} = \left[(\bar{\boldsymbol{x}}_h - \bar{\boldsymbol{x}}_k)^t S^{-1} (\bar{\boldsymbol{x}}_h - \bar{\boldsymbol{x}}_k) \right]^{1/2}$$

변수들 간의 군집을 위해서 사용하는 비유사성은 두 변수 $\boldsymbol{x}_l = (x_{1l,}, \ldots, x_{nl})^t$과 $\boldsymbol{x}_m = (x_{1m,}, \ldots, x_{nm})^t$에 대하여 다음을 사용한다.

$$\text{1-상관계수: } d_{lm} = 1 - r_{lm} = 1 - \frac{s_{lm}}{s_l s_m} \qquad (2.3)$$

여기서 $s_{lm} = \Sigma_{i=1}^{n} (x_{il} - \bar{x}_l)(x_{im} - \bar{x}_m)/(n-1)$은 공분산이고 s_l과 s_m은 각 변수의 표준편차이다. $-1 \le r_{lm} \le 1$은 두 변수의 상관계수이다. 특히, 식 (2.3)에 의해서 크기가 $p \times p$이고 대칭인 비유사성행렬 $D = (d_{lm})$, $l, m = 1, \ldots, p$를 형성한다. 이외에도 Cox와 Cox(2001, Table 1.1)에는 다양한 비유사성계수가 요약되어 있다.

행과 열을 가지는 범주형자료에서 도수(frequency) $f_{rc}(\ge 0)$에 의한 $I \times J$ 분할표를 $F = (f_{rc})$, $r = 1, \ldots, I; c = 1, \ldots, J$ 라고 하면 i와 j번째 두 행범주 사이의 유클리드거리는

다음과 같다.

$$\textbf{행범주 간의 유클리드거리}: d_{ij} = \left[\sum_{c=1}^{J} (f_{ic} - f_{jc})^2 \right]^{1/2}$$

범주형자료의 유클리드거리는 행범주와 더불어 열범주 간에도 동일하게 성립하며 다변량통계기법인 대응분석(correspondence analysis)에서 범주들 간의 기하적인 거리로 정의된다. 따라서 범주들의 군집에 중요한 역할을 한다(최용석, 2001, 2장; 정광모와 최용석, 2009, 12장).

2.1.2 비유사성 측도의 적용

일반적으로 다차원척도법에서는 1.1.2절에서 소개한 [사례 2]의 〈표 1.2〉 우리나라 14개 경제관련기관의 향후 1년간 10가지 경제 전망과 같이 다변량자료행렬 X가 양적(quantative)인 경우 〈표 1.3〉의 개체 간의 유클리드거리를 비유사성으로 하여 다차원척도법을 실시하였다. 물론 앞서 정의한 다른 다양한 거리도 응용될 수 있음은 당연하다. 더불어 양적인 자료 〈표 1.2〉를 0과 1 값을 가진 이진수로 구성된 질적(qualitative) 자료행렬로 변환하여 다차원척도법을 적용하였다. 여기서 경제 전망에 관한 두 기관의 의견이 다른 항목의 수를 전체 항목의 수로 나누어 이를 비유사성 d_{ij}로 정의하였다. 이는 실제 앞서 정의한 두 개체 간의 거리 개념에서 보면 이진수자료의 제곱유클리드거리를 전체 항목수로 나눈 것과 동일하다.

이를 수리적으로 보기 위해 이진수자료행렬 $X=(x_{ir}), i=1,...,n; r=1,...,p$가

$$x_{ir} = \begin{cases} 1 & : \text{ 개체 } i\text{가 } r\text{번째 변수의 성질을 만족한다.} \\ 0 & : \text{ 그렇지 않다.} \end{cases} \tag{2.4}$$

을 만족하는 〈표 1.4〉로부터 한국은행과 동서증권 두 개체만을 고려한 〈표 2.1〉을 생각해 보자. 두 개체 i=한국은행과 j=동서증권에 대해 식 (2.4)에 의해서 다음을 만족한다.

$$(x_{ir} - x_{jr})^2 = \begin{cases} 0 & : x_{ir} = x_{jr} \\ 1 & : x_{ir} \neq x_{jr} \end{cases}$$

〈표 2.1〉 이진수자료의 예

기관	경제 전망									
	1	2	3	4	5	6	7	8	9	10
한국은행	0	1	0	0	0	1	1	1	1	1
동서증권	1	1	0	1	1	0	1	1	1	1

따라서 〈표 2.1〉로부터 10가지 경제 전망에 대한 이들 간의 제곱유클리드거리는

$$\sum_{r=1}^{10}(x_{ir}-x_{jr})^2 = (0-1)^2 + (1-1)^2 + \cdots + (1-1)^2 = 4$$

이다. 이는 경제 전망에 관한 두 기관의 의견이 다른 항목의 수이다. 이 값을 전체 항목의 개수 10으로 나누어 주면 0.4가 된다. 이는 이미 1.1.2절의 [사례 2]에서 두 개체 $i=$한국은행과 $j=$동서증권에 대한 비유사성 $d_{ij}=0.4$와 동일하다.

이외에도 여러 가지 비유사성을 나타낼 거리들에 대해서는 Mardia 외 2인(1979, pp. 375-381)과 Cox와 Cox(2001, Table 1.1)를 참고하기를 바란다. 특히, 식 (2.2)의 가중유클리드거리를 고려하는 다차원척도법을 1.2절에서 Carroll과 Chang(1970)의 INDSCAL이라 하였다. 이를 수정하여 SAS에서는 PROC ALSCAL에서 이를 수용하였고 PROC MDS 문에서는 옵션 〈COEF=DIAGONAL〉이 그 역할을 대신하고 있다. 유클리드거리에 의한 다차원척도법은 옵션 〈COEF=IDENTITY〉(디폴트)이 제공한다. 이들에 대한 자세한 내용은 4장의 삼원 다차원척도법을 참고하기를 바란다.

2.1.3 비유사성 측도의 계산을 위한 프로그램

이 절에서는 1.1.2절의 [사례 2]의 14개 경제관련기관의 10가지 경제 전망〈표 1.2〉에 대하여 비유사성 측도 중 유클리드거리를 계산하고 이를 비유사성행렬자료로 하여 다차원척도법을 위한 SAS/PROC MDS 프로그램을 소개하려 한다.

[프로그램 2.1]은 〈표 1.2〉와 같이 원자료가 양적인 다변량자료행렬 X를 이루며 이로부터 유클리드거리 등 다양한 거리와 같은 비유사성 측도 계산을 행렬프로그래밍언어인 SAS/IML을 활용하여 선보이고 있다. 이를 통하여 개체 간의 유클리드거리로 이루어진 〈표 1.3〉과 같은 비유사성행렬자료를 얻게 되고 더불어 SAS/PROC MDS와 연동하여 그 결과인 비유사성행렬자료에 대해 직접 다차원척도법을 실시할 수 있다. [결과 2.1]-① Euclidean distance에서 유클리드거리를 보여 주고 있다. 물론, 1.4.2절의 [프로그램 1.2]-(a)는 〈표

1.3〉의 비유사성행렬자료가 입력자료로 제공되어 있을 때 다차원척도법 적용관점에서 작성된 것이며 실행결과는 [결과 1.2]-(a)에서 자세히 살펴보았다. 비록 [프로그램 2.1]의 실행결과는 [프로그램 1.2]-(a)의 결과와 동일하나 종합적으로 요약차원에서 [결과 2.1]에 다시 정리하고 있다.

[프로그램 2.1] 원자료로부터 다차원척도법을 위한 SAS/IML과 MDS

```
/* Creating dissimilarity matrix from a raw data*/
OPTIONS PS=150 LS=250;

data economy;
      input id $ x1-x10;
lines;
hankook     8.2   1950   698   650   98   280   630   3.0   5.7   12.0
daewoo1     9.5   2100   710   620   110  270   640   2.7   4.5   12.0
dongseo1    9.0   2000   690   630   100  290   630   2.6   6.0   12.0
kyungrung   7.8   1850   668   660   88   280   620   3.0   6.4   13.8
sangkon     8.5   1928   710   670   90   290   620   3.0   6.0   10.0
choongso    9.0   1958   710   615   95   280   603   4.0   6.0   10.0
hyundai     8.5   1900   700   610   100  250   620   3.2   5.5   14.0
samsung     8.0   1900   700   640   100  280   640   2.7   5.5   10.0
sunkyung    8.5   1950   700   620   120  300   630   2.7   7.0   12.0
daewoo2     7.9   1900   697   645   95   280   610   2.9   6.8   15.0
shinhan     9.0   2030   700   620   100  275   630   3.0   7.0   13.0
dongseo2    8.5   1950   690   630   90   290   630   2.9   6.0   12.0
korea       9.9   1870   729   649   123  260   616   3.4   6.1   15.8
chungang    8.5   1700   700   630   90   260   620   4.0   6.0   14.0
;
proc iml;
   reset nolog;
   start dissim;

   /* Creating a matrix from a SAS data set */
   use economy;
   read all var{id} into level;
   read all var{x1 x2 x3 x4 x5 x6 x7 x8 x9 x10} into X; /* choose variables */

   n=nrow(X);
   p=ncol(X);

 /* standardization */
       Xbar=X[+,]/n;
       Y=X-J(n,1,1)*Xbar;
       variance=Y`*Y/(n-1);

       if std=1 then
                do;
                   s=diag(variance);
                   Y=Y*sqrt(inv(s));
                end;
       else Y=Y;
```

```
   D_Eu=shape(.,n,n);
   D_Ma=shape(.,n,n);
   D_City=shape(.,n,n);

   /* Generating Dissimilarity Matrix */
   do i=1 to n;
      do j=1 to n;
         A=Y[i,];
         B=Y[j,];
         C=A-B;
         absC=ABS(C);

         Eu=C*C`; /* Euclidean Distance */
         D_Eu[i,j]=sqrt(Eu);

         Ma=C*inv(variance)*C`; /* Mahalanobis Distance */
         D_Ma[i,j]=sqrt(Ma);

         City=absC[+];
         D_City[i,j]=City; /* City-block distance */
      end;
   end;

    print D_Eu[rowname=level format=5.3];
   /* print D_Ma[rowname=level format=5.3];*/
   /* print D_City[rowname=level format=5.3];*/

  /* Creating SAS Data Set for PROC MDS */
   create Euclid from D_Eu[rowname=level];
   append from D_Eu[rowname=level];
   close Euclid;

   create Mahala from D_Ma[rowname=level];
   append from D_Ma[rowname=level];
   close Mahala;

   create City from D_City[rowname=level];
   append from D_City[rowname=level];
   close City;

   Finish;

std=1;
   run dissim;
quit;

ods graphics on;
proc mds data=Euclid pfinal;
        id level;
```

[결과 2.1]-②는 비계량형 다차원척도법의 $ST1$(크루스칼의 스트레스 1)의 최소화 알고리즘을 위한 수치적 계산과정과 2차원 다차원척도그림을 위한 형상좌표를 제공한다. 결과 ③은 결과 ②의 마지막 반복의 〈Badness-of-Fit Criterion〉값 0.1496을 제공하고 있다. 이는 약 0.15이고 크루스칼의 판별 기준인 〈표 1.9〉에 따르면 〈0.10 보통(fair)〉보다는 낮고

〈0.20 나쁨(poor)〉보다는 높은 정도를 나타낸다. 결과 ④는 〈그림 1.2〉 경제관련기관의 경제 전망의 다차원척도그림으로 그 해석도와 함께 이미 1.1.2절에서 살펴보았다.

[결과 2.1] [프로그램 2.1]의 수행결과

①

Euclidean distance

D_Eu														
hankook	0.000	4.166	2.293	3.269	2.390	4.410	3.573	1.846	3.700	3.131	3.118	1.784	5.362	4.343
daewoo1	4.166	0.000	3.672	6.925	5.476	5.723	4.432	4.197	5.189	6.345	4.344	4.443	5.571	6.693
dongseo1	2.293	3.672	0.000	4.110	3.496	4.635	4.084	2.839	2.854	3.800	2.363	1.515	5.684	5.425
kyungrun	3.269	6.925	4.110	0.000	4.157	5.837	4.812	4.322	5.207	2.789	4.653	3.291	6.926	4.474
sangkong	2.390	5.476	3.496	4.157	0.000	4.316	5.177	3.182	4.631	3.820	4.265	3.061	5.856	5.124
choongso	4.410	5.723	4.635	5.837	4.316	0.000	4.231	5.220	5.231	4.796	4.260	4.263	5.563	4.407
hyundai	3.573	4.432	4.084	4.812	5.177	4.231	0.000	4.233	5.088	3.928	3.592	3.820	4.699	3.336
samsung	1.846	4.197	2.839	4.322	3.182	5.220	4.233	0.000	3.880	4.468	3.920	2.469	6.336	5.114
sunkyung	3.700	5.189	2.854	5.207	4.631	5.231	5.088	3.880	0.000	4.159	2.990	3.401	5.696	6.064
daewoo2	3.131	6.345	3.800	2.789	3.820	4.796	3.928	4.468	4.159	0.000	3.549	3.278	5.349	4.192
shinhan	3.118	4.344	2.363	4.653	4.265	4.260	3.592	3.920	2.990	3.549	0.000	2.648	4.994	4.934
dongseo2	1.784	4.443	1.515	3.291	3.061	4.263	3.820	2.469	3.401	3.278	2.648	0.000	6.088	4.568
korea	5.362	5.571	5.684	6.926	5.856	5.563	4.699	6.336	5.696	5.349	4.994	6.088	0.000	5.176
chungang	4.343	6.693	5.425	4.474	5.124	4.407	3.336	5.114	6.064	4.192	4.934	4.568	5.176	0.000

②

Shape=TRIANGLE Condition=MATRIX Level=ORDINAL

Coef=IDENTITY Dimension=2 Formula=1 Fit=1
Mconverge=0.01 Gconverge=0.01 Maxiter=100 Over=2 Ridge=0.0001

Iteration	Type	Badness-of-Fit Criterion	Change in Criterion	Convergence Measure	
				Monotone	Gradient
0	Initial	0.2776	.	.	.
1	Monotone	0.2049	0.0727	0.1760	0.5150
2	Gau-New	0.1708	0.0341	.	.
3	Monotone	0.1656	0.005190	0.0336	0.2572
4	Gau-New	0.1598	0.005814	.	.
5	Monotone	0.1532	0.006609	0.0436	0.1787
6	Gau-New	0.1525	0.000733	.	.
7	Monotone	0.1513	0.001152	0.0174	0.1406
8	Gau-New	0.1508	0.000518	.	.
9	Monotone	0.1503	0.000486	0.0117	0.0958
10	Gau-New	0.1501	0.000215		.
11	Monotone	0.1500	0.000130	0.006270	0.0726
12	Gau-New	0.1496	0.000379		0.006127

Configuration		
	Dim1	Dim2
hankook	0.40	0.06
daewoo1	-1.11	1.75
dongseo1	0.21	0.62
kyungrun	1.59	-0.78
sangkong	1.22	0.05
choongso	-0.88	-1.33
hyundai	-0.84	-0.32
samsung	0.81	0.74
sunkyung	0.40	1.45
daewoo2	0.67	-0.77
shinhan	-0.26	0.47
dongseo2	0.42	0.23
korea	-2.54	-0.28
chungang	-0.09	-1.89

③

MATRIX	Number of Nonmissing Data	Weight	Badness-of-Fit Criterion	Distance Correlation	Uncorrected Distance Correlation
1	91	1.00	0.15	0.94	0.99

④ 〈그림 1.2〉 경제관련기관의 경제 전망의 비계량형 다차원척도그림:
(a) 원자료의 유클리드거리

2.2 유사성

이번에는 비유사성과는 정반대의 개념인 두 개체 i와 j의 유사성 c_{ij}를 나타낼 측도로는 유사성계수(similarity coefficient)가 있다. 대개 유사성은 2.1절의 비유사성 조건 1)-조건 3)과 같은 성질을 갖고 있다(Mardia 외 2인, 1979, p. 382; du Toit 외 2인, 1986, p. 76). 다만 조건 2)에서는 $c_{ii}=1$을 만족한다. 특히, 이진수자료인 경우 개체 간의 유사성 측도와 변수들 간의 유사성 측도에 대해서 알아보고자 한다.

2.2.1 유사성 측도

자료행렬 $X=(x_{ir}), i=1,\ldots,n; r=1,\ldots,p$가 식 (2.4)를 만족하는 이진수자료인 경우 먼저 두 개체 간의 유사성을 나타내는 몇 가지 유사성계수를 알아보자. 이를 위해 〈표 2.2〉의 2×2 조합표(association table)를 생각해 보자.

이는 두 개체 i와 j가 0과 1을 가지는 이진수자료로부터 새로 구성된 이원 분할표 형태로 a는 1-1 짝의 수, b는 1-0 짝의 수 등을 나타내며 따라서 $p=a+b+c+d$는 이들의 전체 합이 된다.

〈표 2.2〉 이진수자료의 두 개체에 대한 2×2 조합표

		개체 j		
		1	0	합계
개체 i	1	a	b	$a+b$
	0	c	d	$c+d$
합계		$a+c$	$b+d$	p

Johnson과 Wichern(2002, pp. 670-679)은 〈표 2.2〉로부터 개체 i와 j의 유사성 c_{ij}를 나타내는 유사성계수 8가지를 소개하고 있고 이를 인용하여 〈표 2.3〉에 정리하였다. 이외에도 Cox와 Cox(2001, Table 1.2)에는 이진수자료에 대한 유사성계수를 더 제공하고 있다. 이들 중에서 가장 보편적으로 이용되는 것은 ①, ④, ⑤ 세 가지로 이를 다음과 같이 부른다.

- 단순매칭계수(simple matching coefficient): $c_{ij}=(a+d)/p$.
- 러셀-라오계수(Russell and Rao coefficient): $c_{ij}=a/p$.
- 자카드계수(Jaccard's coefficient): $c_{ij}=a/(a+b+c)$.

이러한 계수들에 의한 크기가 $n \times n$인 유사성행렬을 각각 C_S, C_R, C_J라고 하자. 더군다나, C_R과 C_J 그리고 C_S는 이진수 자료행렬 $X=(x_{ir}), i=1,...,n; r=1,...,p$로부터 식 (2.5)와 식 (2.6)의 행렬연산을 이용하면 쉽게 얻어진다(Mardia 외 2인, 1979, p. 383).

$$C_R = \frac{XX^t}{p} \text{와 } C_J = \frac{XX^t}{p-d}. \tag{2.5}$$

$$C_S = \frac{[XX^t + (J-X)(J-X)^t]}{p} \tag{2.6}$$

여기서 $J=1_n 1_n^t$는 크기가 $n \times n$이고 모든 원소가 1인 행렬이고 C_S의 대각 원소는 모두 1이다. 그리고 C_R, C_J, C_S는 양반정치행렬이다.

〈표 2.3〉 이진수자료의 유사성계수

유사성계수	참고 사항
① $\frac{a+d}{p}$	1-1짝과 0-0짝에 대해 같은 가중치
② $\frac{2(a+d)}{2(a+d)+b+c}$	1-1짝과 0-0짝에 대해 두 배 가중치
③ $\frac{a+d}{a+d+2(b+c)}$	불일치 짝에 두 배 가중치
④ $\frac{a}{p}$	분자에 1-1짝만 있음
⑤ $\frac{a}{a+b+c}$	분자 또는 분모에 0-0짝이 없음
⑥ $\frac{2a}{2a+b+c}$	분자 또는 분모에 0-0짝이 없음. 1-1짝에 두 배 가중치
⑦ $\frac{a}{a+2(b+c)}$	분자 또는 분모에 0-0짝이 없음. 불일치 짝에 두 배 가중치
⑧ $\frac{a}{b+c}$	0-0짝을 제외하고 불일치 짝에 대한 일치 짝의 비

이번에는 이진수자료행렬 $X=(x_{ir}), i=1,...,n; r=1,...,p$로부터 〈표 2.2〉의 두 개체에 대한 2×2 조합표 대신에 두 변수에 대한 2×2 조합표를 구성할 수 있다. 이 경우 조

합표는 분할표(contingency table)가 되며 〈표 2.4〉와 같고 여기서 $n = n_{11} + n_{12} + n_{21} + n_{22}$은 총 개체의 수이다.

〈표 2.4〉 이진수자료에서 두 변수에 대한 2×2 분할표

		변수 m		합계
		1	0	
변수 l	1	n_{11}	n_{12}	$n_{11} + n_{12}$
	0	n_{21}	n_{22}	$n_{21} + n_{22}$
합계		$n_{11} + n_{21}$	$n_{12} + n_{22}$	n

〈표 2.4〉로부터 두 변수의 곱적률 상관계수(product moment correlation coefficient)는

$$r = \frac{n_{11}n_{22} - n_{12}n_{21}}{[(n_{11} + n_{12})(n_{21} + n_{22})(n_{11+}n_{21})(n_{12} + n_{22})]^{1/2}} \tag{2.7}$$

가 되며, 이 값을 두 변수간의 유사성 측도로 사용할 수 있다. 식 (2.7)의 곱적률 상관계수는 귀무가설로 두 개 범주형 변수들 간의 유사성이 없다는 독립성을 검정하는 카이제곱 검정통계량 χ^2과는 수리적으로 $\chi^2 = nr^2$의 관계가 성립한다. n이 고정된 경우 유사성을 나타내는 곱적률 상관계수가 크면 귀무가설을 기각하게 되고 독립성에서 벗어나 통계적으로 유의하게 유사성이 있음을 말하게 된다.

참고로 개체들 간의 유사성계수를 나타낸 〈표 2.3〉에서 p대신 n을 사용하고 $a = n_{11}$, $b = n_{12}$, $c = n_{21}$, $d = n_{22}$로 대체하면 변수들 간의 유사성 측도를 만들 수 있다.

지금까지 이진수자료에 대한 개체들 간의 유사성 측도와 변수들 간의 유사성 측도에 대해 살펴보았다. 일반적으로 양적자료행렬 $X = (x_{ir}), i = 1,...,n; r = 1,...,p$에서는 두 변수 $\boldsymbol{x}_l = (x_{1l},...,x_{nl})^t$과 $\boldsymbol{x}_m = (x_{1m},...,x_{nm})^t$에 대한 유사성 측도로 식 (2.3)에서 사용된 피어슨의 상관계수

$$c_{lm} = r_{lm} = s_{lm}/s_l s_m \tag{2.8}$$

를 이용함은 잘 알려져 있다. 특히, 음의 상관계수는 때때로 그 절댓값을 유사성으로 사용

한다. 식 (2.8)에서 정의한 유사성에 의해 크기가 $p \times p$이고 대각원소가 1인 대칭인 유사성 행렬 $C = (c_{lm})$, $l, m = 1, \ldots, p$을 형성할 수 있다.

끝으로 양적변수와 질적변수가 혼합되어 있는 경우 다음과 같은 유사성계수

$$c_{ij} = 1 - \frac{1}{p} \sum_{r=1}^{p} w_r |x_{ir} - x_{jr}|,$$

를 정의할 수 있으며 이로부터 유사성행렬은 양반정치이며 여기서 만약에 l번째가 질적변수이면 $w_l = 1$이고, 양적변수이면 $w_l = 1/R_l$이며 $R_l = \max_{ij} |x_{il} - x_{jl}|$은 l번째 변수의 범위(range)라고 한다(Mardia 외 2인, 1979, pp. 383-384).

2.2.2 유사성 측도의 적용

이 절에서는 〈표 2.3〉에서 정의한 보편적인 유사성계수 중 단순매칭계수를 계산하는 사례를 위하여 〈표 2.1〉로부터 두 개체 i=한국은행과 j=동서증권에 대한 〈표 2.5〉와 같은 2×2 조합표를 구성하자.

이는 경제 전망이 같은 짝 1-1과 0-0이 각각 5개와 1개이고, 나머지 4개는 경제 전망이 다른 짝 1-0 혹은 0-1로 이루어져 있다. 따라서 단순매칭계수를 활용한 두 개체 i=한국은행과 j=동서증권에 대한 유사성은 $c_{ij} = (a+d)/p = (5+1)/10 = 0.6$이 된다.

〈표 2.5〉 한국은행과 동서증권의 2×2 조합표

		동서증권		합계
		1	0	
한국은행	1	5	1	6
	0	3	1	4
합계		8	2	10

같은 사례에 대하여 2.1.2절에서 두 개체 간의 제곱유클리드거리로 살펴본 비유사성은 $d_{ij} = 0.4$로, 〈표 2.5〉에 따르면 경제 전망이 다른 짝 1-0 혹은 0-1의 수 $b+c = 1+3 = 4$에 전체 합 $p = 10$으로 나눈 $(b+c)/p = 4/10$의 결과이다. 따라서 이들 유사성 c_{ij}와 비유사성 d_{ij}의 관계는 $d_{ij} = 1 - c_{ij}$이 됨을 알 수 있다. 이와 같이 유사성과 비유사성 간의

다양한 변환은 2.3절에서 자세히 설명하고 있다.

다음으로 변수들 간의 유사성 측도를 구하고 이를 평가하는 예를 살펴보자. 〈표 1.4〉의 우리나라 14개 경제관련기관의 10가지 경제 전망 중에 1=성장률과 2=GNP에 대해서 〈표 2.6〉과 같이 2×2 분할표를 구성할 수 있다.

〈표 2.6〉 1=성장률과 2=GNP의 2×2 분할표

		2=GNP		합계
		1	0	
1=성장률	1	6	4	10
	0	1	3	4
합계		7	7	14

〈표 2.6〉으로부터 식 (2.7)의 곱적률 상관계수를 계산하면

$$r = \frac{6\times3-4\times1}{[10\times4\times7\times7]^{1/2}} = \frac{\sqrt{10}}{10} \simeq 0.316$$

이다. 이 값은 낮은 상관을 보여 성장률과 GNP는 유사하지 못함을 보여 준다. 더불어 이 곱적률 상관계수를 가지고 카이제곱통계량을 계산하여 보면 $\chi^2 = n\times r^2 = 14\times(0.316)^2 \simeq 1.4$가 얻어진다.

실제로 이 값은 〈표 2.6〉으로부터 〈귀무가설 H_0 : 성장률과 GNP는 연관성이 없다.〉라는 범주형자료분석의 독립성 검정을 위한 카이제곱 검정통계량

$$\chi^2 = \left(\frac{6-5}{\sqrt{5}}\right)^2 + \left(\frac{4-5}{\sqrt{5}}\right)^2 + \left(\frac{1-2}{\sqrt{2}}\right)^2 + \left(\frac{3-2}{\sqrt{2}}\right)^2 = 1.4$$

와 동일하며 자유도는 $df = (2-1)(2-1) = 1$이다. 이 경우 유의수준 5%에서 기각값은 3.84로 검정통계량값 1.4보다 크므로 귀무가설을 기각하지 못한다. 즉, 성장률과 GNP는 연관성이 없으며 통계적으로 유사하지 못하다.

2.3 유사성과 비유사성의 변환

일반적으로 다변량자료행렬 X로부터 두 개체 i와 j간의 거리로 나타낸 비유사성 d_{ij}가 크면 두 개체 간의 유사성 c_{ij}는 작아지며, 반대로 비유사성이 작아지면 유사성은 크게 된다. 비유사성과 유사성의 사이의 가장 큰 차이점은 거리를 나타내는 비유사성은 양의 값을 나타내는 반면에 유사성은 0과 1 사이의 값을 나타낸다. 가장 잘 알려진 비유사성에서 유사성으로의 변환(transformation)은

$$c_{ij} = \frac{1}{1 + d_{ij}}$$

이며 $0 < c_{ij} \leq 1$을 만족한다. 그 반대의 경우는 비유사성의 성질인 삼각 부등식 때문에 명백하지 않다. Mardia 외 2인(1979, p. 402)과 du Toit 외 2인(1986, p. 77)에 따르면 만약에 유사성행렬 $C = (c_{ij})$가 양반정치(positive semidefinite)를 만족하는 경우

$$d_{ij} = \sqrt{c_{ij} - 2c_{ij} + c_{ij}} \tag{2.9}$$

를 표준변환이라고 하며 이로부터 얻어지는 비유사성행렬 $D = (d_{ij})$는 유클리드거리행렬이 된다.

단순매칭계수를 유사성으로 사용할 경우에는 2.2.2절에서 살펴보았지만 변환 $d_{ij} = 1 - c_{ij}$도 고려할 수 있다. 특히, $c_{ii} = c_{jj} = 1$인 경우는

$$d_{ij} = \sqrt{2(1 - c_{ij})}$$

로 단순하게 변환된다.

끝으로 [프로그램 2.2]와 [결과 2.2]는 1.1.2절의 [사례 2]의 우리나라 14개 경제관련기관의 10가지 경제 전망에 대한 이진수자료 〈표 1.4〉로부터 유사성 측도 중 단순매칭계수를 유사성 c_{ij}로 하고 변환 $d_{ij} = 1 - c_{ij}$를 통해 얻어지는 비유사성행렬자료의 다차원척도법을 위한 것이다. SAS/IML 프로그램에서 식 (2.5)와 식 (2.6)의 유사성행렬에 관한 식을 이용하고 있다. SAS/IML의 문법에 대해선 SAS Institute, Inc.(2004)를 참고하기 바란다.

특히, 이진수자료에서 비교하고자 하는 개체 수가 매우 많은 경우 [프로그램 2.2]의 유용성은 높다고 여겨진다. 실제 SAS/IML과 PROC MDS가 결합되어 작성된 프로그램의 알고리즘은 2.2.1절의 유사성 측도 중 단순매칭계수와 변환이 가능하므로 이를 통하여 비유사성행렬자료를 제공하게 된다. 실제로 [결과 2.2]-① Similarity matrix of simple matching coefficient와 Dissimilarity matrix가 이를 보여 주고 있다. 나머지는 1.4.2절의 [결과 1.2]-(b)와 동일하고 해석도 자세히 언급되어 있어 생략한다.

[프로그램 2.2] 이진수자료로부터 다차원척도법을 위한 프로그램

```
        data binary;
          input id $10. (v1 - v10) (1.) ;
        cards;
        hankook      0100011111
        daewoo1      1111111111
        dongseo1     1101101111
        kyungrun     0000010100
        sangkong     1010000111
        choongso     1111010011
        hyundai      1011110010
        samsung      0010111111
        sunkyung     1111101101
        daewoo2      0000010100
        shinhan      1111111100
        dongseo2     1101001111
        korea        1010110000
        chungang     1011010010
        ;

proc iml;
   reset nolog;
   start dissim;

   /* Creating a matrix from a SAS data set */
   use economy_binary;
   read all var{id} into level;
   read all var{x1 x2 x3 x4 x5 x6 x7 x8 x9 x10} into X;

   n=nrow(X);
   p=ncol(X);
   J= J(n, n, 1);

   Cr = X*X`/p; /* Russell-Rao coefficient*/
   Cs = Cr + (J-X)*(J-X)` /p; /* Simple matching coefficient */
   Ds = 1- Cs; /* Dissimilarity matrix */
```

```
  print, "Similarity matrix of simple matching coefficient",
                        Cs[rowname=level format=3.1];
  print, "Dissimilarity matrix", Ds[rowname=level format=3.1];

     /* Creating SAS Data Set for PROC MDS */
   create DM from Ds[rowname=level];
   append from Ds[rowname=level];
   close DM;

       Finish;
    run dissim;
 quit;

ods graphics on;
proc mds data=DM pfinal;
       id level;
```

[결과 2.2] [프로그램 2.2]의 수행결과

①

Similarity matrix of simple matching coefficient

Cs														
hankook	1.0	0.6	0.6	0.6	0.5	0.5	0.2	0.7	0.4	0.6	0.4	0.7	0.2	0.3
daewoo1	0.6	1.0	0.8	0.2	0.5	0.7	0.6	0.7	0.8	0.2	0.8	0.7	0.4	0.5
dongseo1	0.6	0.8	1.0	0.2	0.5	0.5	0.4	0.5	0.8	0.2	0.6	0.9	0.2	0.3
kyungrun	0.6	0.2	0.2	1.0	0.5	0.3	0.4	0.5	0.2	1.0	0.4	0.3	0.6	0.5
sangkong	0.5	0.5	0.5	0.5	1.0	0.6	0.5	0.6	0.5	0.5	0.3	0.6	0.5	0.6
choongso	0.5	0.7	0.5	0.3	0.6	1.0	0.7	0.4	0.5	0.3	0.5	0.6	0.5	0.8
hyundai	0.2	0.6	0.4	0.4	0.5	0.7	1.0	0.5	0.4	0.4	0.6	0.3	0.8	0.9
samsung	0.7	0.7	0.5	0.5	0.6	0.4	0.5	1.0	0.5	0.5	0.5	0.4	0.5	0.4
sunkyung	0.4	0.8	0.8	0.2	0.5	0.5	0.4	0.5	1.0	0.2	0.8	0.7	0.4	0.3
daewoo2	0.6	0.2	0.2	1.0	0.5	0.3	0.4	0.5	0.2	1.0	0.4	0.3	0.6	0.5
shinhan	0.4	0.8	0.6	0.4	0.3	0.5	0.6	0.5	0.8	0.4	1.0	0.5	0.6	0.5
dongseo2	0.7	0.7	0.9	0.3	0.6	0.6	0.3	0.4	0.7	0.3	0.5	1.0	0.1	0.4
korea	0.2	0.4	0.2	0.6	0.5	0.5	0.8	0.5	0.4	0.6	0.6	0.1	1.0	0.7
chungang	0.3	0.5	0.3	0.5	0.6	0.8	0.9	0.4	0.3	0.5	0.5	0.4	0.7	1.0

Dissimilarity matrix

Cs														
hankook	0.0	0.4	0.4	0.4	0.5	0.5	0.8	0.3	0.6	0.4	0.6	0.3	0.8	0.7
daewoo1	0.4	0.0	0.2	0.8	0.5	0.3	0.4	0.3	0.2	0.8	0.2	0.3	0.6	0.5
dongseo1	0.4	0.2	0.0	0.8	0.5	0.5	0.6	0.5	0.2	0.8	0.4	0.1	0.8	0.7
kyungrun	0.4	0.8	0.8	0.0	0.5	0.7	0.6	0.5	0.8	0.0	0.6	0.7	0.4	0.5
sangkong	0.5	0.5	0.5	0.5	0.0	0.4	0.5	0.4	0.5	0.5	0.7	0.4	0.5	0.4
choongso	0.5	0.3	0.5	0.7	0.4	0.0	0.3	0.6	0.5	0.7	0.5	0.4	0.5	0.2
hyundai	0.8	0.4	0.6	0.6	0.5	0.3	0.0	0.5	0.6	0.6	0.4	0.7	0.2	0.1
samsung	0.3	0.3	0.5	0.5	0.4	0.6	0.5	0.0	0.5	0.5	0.5	0.6	0.5	0.6
sunkyung	0.6	0.2	0.2	0.8	0.5	0.5	0.6	0.5	0.0	0.8	0.2	0.3	0.6	0.7
daewoo2	0.4	0.8	0.8	0.0	0.5	0.7	0.6	0.5	0.8	0.0	0.6	0.7	0.4	0.5
shinhan	0.6	0.2	0.4	0.6	0.7	0.5	0.4	0.5	0.2	0.6	0.0	0.5	0.4	0.5
dongseo2	0.3	0.3	0.1	0.7	0.4	0.4	0.7	0.6	0.3	0.7	0.5	0.0	0.9	0.6
korea	0.8	0.6	0.8	0.4	0.5	0.5	0.2	0.5	0.6	0.4	0.4	0.9	0.0	0.3
chungang	0.7	0.5	0.7	0.5	0.4	0.2	0.1	0.6	0.7	0.5	0.5	0.6	0.3	0.0

②

Shape=TRIANGLE Condition=MATRIX Level=ORDINAL
Coef=IDENTITY Dimension=2 Formula=1 Fit=1
Mconverge=0.01 Gconverge=0.01 Maxiter=100 Over=2 Ridge=0.0001

Iteration	Type	Badness-of-Fit Criterion	Change in Criterion	Convergence Measure	
				Monotone	Gradient
0	Initial	0.2275	.	.	.
1	Monotone	0.2131	0.0144	0.0751	0.5269
2	Gau-New	0.1769	0.0362	.	.
3	Monotone	0.1763	0.000561	0.0339	0.2793
4	Gau-New	0.1703	0.006019	.	.
5	Monotone	0.1670	0.003284	0.0366	0.1890
6	Gau-New	0.1668	0.000199	.	.
7	Monotone	0.1667	0.0000930	0.008599	0.1841
8	Gau-New	0.1639	0.002850	.	0.007253

Configuration		
	Dim1	Dim2
hankook	0.34	1.55
daewoo1	0.97	-0.25
dongseo1	1.50	0.37
kyungrun	-1.61	0.92
sangkong	-0.48	0.34
choongso	0.09	-1.04
hyundai	-0.70	-1.28
samsung	-0.06	0.99
sunkyung	1.49	-0.32
daewoo2	-1.61	0.92
shinhan	0.92	-0.92
dongseo2	1.36	0.63
korea	-1.41	-0.83
chungang	-0.80	-1.08

③

MATRIX	Number of Nonmissing Data	Weight	Badness-of-Fit Criterion	Distance Correlation	Uncorrected Distance Correlation
1	91	1.00	0.16	0.91	0.99

④

〈그림 1.2〉 경제관련기관의 경제 전망의 비계량형 다차원척도그림:
(b) 이진수자료의 제곱유클리드거리

2.4 비대칭 비유사성과 유사성의 변환

지금까지 일반적으로 다변량자료행렬 X로부터 두 개체 i와 j 간의 비유사성 d_{ij} 또는 유사성 c_{ij}에 의한 행렬자료는 2.1.1절에서 대칭성 조건 $d_{ij} = d_{ji}$ (또는 $c_{ij} = c_{ji}$)을 만족하였다. 그러나 경우에 따라 개체들 간에 대칭성이 만족되지 못하는 비대칭성(asymmetric)

$$d_{ij} \neq d_{ji} \text{ (또는 } c_{ij} \neq c_{ji}) \tag{2.10}$$

이 발생하기도 한다.

〈표 2.7〉 아라비아 숫자에 대한 모스부호의 비대칭 유사성자료

1.----	84	63	13	8	10	8	19	32	57	55
2..---	62	89	53	20	5	14	20	21	16	11
3...--	16	59	86	31	23	41	16	17	08	10
4....-	6	23	38	89	42	44	32	10	3	3
5.....	12	8	27	56	90	42	24	10	6	5
6-....	12	14	33	34	30	86	69	14	5	14
7--...	20	25	17	24	18	65	85	70	20	13
8---..	37	25	16	13	10	22	65	88	61	26
9----.	57	28	9	7	5	8	31	58	91	78
0-----	52	18	9	7	5	18	15	39	79	94

〈표 2.7〉은 Everitt와 Dunn(1991, Table 5.8), Borg와 Groenen(1997, Table 4.2)의 26자 알파벳(A~Z)과 10개의 아리비아 숫자(0~9)에 대한 모스부호(Morse code)에서 아리비아 숫자에 대한 것만 정리한 자료이다. 이는 모스부호를 모르는 598명에게 청각적으로 기계에 의해서 두 수의 모스부호 신호를 짝을 지어 두 번(1 다음 2, 2 다음 1과 같이) 들려주고 두 신호의 짝이 같다고 생각하는 사람들의 비율을 나타낸 것이다. 〈표 2.7〉은 대표적으로 식 (2.10)을 만족하지 않는 비대칭 유사성행렬이다.

이와 같이 비대칭 유사성(또는 비유사성)을 대칭성이 만족되도록 만드는 가장 대표적인 두 변환을 비유사성의 경우로 예를 들면

$$d_{ij}^{M} = \frac{d_{ij} + d_{ji}}{2}, \ d_{ij}^{N} = \frac{d_{ij} - d_{ji}}{2} \tag{2.11}$$

이다. 식 (2.11)에서 새로운 비유사성 d_{ij}^{M}은 대칭성 조건을 만족하며, d_{ij}^{N}은 반(skew)대칭성 조건 $d_{ij}^{N} = -d_{ij}^{N}$을 만족한다.

Borg와 Groenen(1997, pp. 402-407) 그리고 Cox와 Cox(2001, pp. 116-121)에 따르면 일반적으로 정방행렬은 대칭행렬과 반대칭행렬로 이루어진다. 따라서 비대칭 비유사성 행렬 $D = (d_{ij})$도 정방행렬로 각각 다음과 같이 분해된다.

$$D = D_M + D_N \tag{2.12}$$

여기서 $D_M = (d_{ij}^M)$과 $D_N = (d_{ij}^N)$은 식 (2.11)로부터 이루어진다. 특히, 반대칭행렬 D_N의 대각원소는 항상 0이다.

식 (2.12)에서 대칭성을 만족하는 $D_M = (d_{ij}^M)$에 대하여 기존의 다차원척도법을 적용하게 되며 반대칭성을 만족하는 $D_N = (d_{ij}^N)$에 대해서는 $D_N^t D_N$의 고유값 $\lambda_1 = \lambda_2 > \lambda_3 = \lambda_4 > \cdots > \lambda_{n/2} = \lambda_{n/2} > 0$에서 같은 값을 가지는 첫 번째 두 고유값 λ_1, λ_2에 대하여 크기가 $n \times 2$ 고유벡터 쌍 $(\boldsymbol{v}_1, \boldsymbol{v}_2)$에 대한 2차원상의 그림 통해 개체들의 관계를 설명할 수 있다.

Borg와 Groenen(1997, pp. 402-404)에서는 만약에 식 (2.12)의 행렬분해에서 행렬 D와 D_M의 대각원소의 값이 같고 반대칭행렬 D_N의 대각원소는 전부 0의 값을 가지므로 대각원소를 제외하고 다음과 같이 원소제곱합의 분해를 고려한다.

$$\sum_{i,j}^{n} d_{ij}^2 = \sum_{i,j}^{n} d_{ij}^{M^2} + \sum_{i,j}^{n} d_{ij}^{N^2} . \tag{2.13}$$

그리고 식 (2.13)과 대응하여 이를 다음과 같이 나타내자.

$$SSD = SSD_M + SSD_N. \tag{2.14}$$

전체 제곱합 SSD에서 차지하는 SSD_M와 SSD_N의 상대적 비율

$$RSSD_M = \frac{SSD_M}{SSD} \times 100\%,$$
$$RSSD_N = \frac{SSD_N}{SSD} \times 100\% \tag{2.15}$$

을 고려하면 식 (2.15)에서 $RSSD_M$이 $RSSD_N$에 비해 상대적으로 매우 크다면 이는 식 (2.12)의 비대칭 비유사성행렬 $D = (d_{ij})$에서 대칭행렬 $D_M = (d_{ij}^M)$이 우세하며 반대칭행렬 $D_N = (d_{ij}^N)$의 역할이 미미함을 나타내지만 개체 간의 흥미 있는 부분은 보여질 수는 있다. 그러나 이와 같은 경우 일반적으로 대칭성을 만족하는 $D_M = (d_{ij}^M)$에 대하여 기존의 다차

원척도법을 적용하면 된다. 더불어 비대칭 유사성행렬 $C=(c_{ij})$에 대해서도 비대칭 비유사성행렬 $D=(d_{ij})$과 같이 식 (2.11) ~ 식 (2.15)의 동일한 수리적 개념과 절차를 적용할 수 있다.

Cox와 Cox(2001, pp. 116-121)과 Borg와 Groenen(1997, pp. 402-407)에서는 이외에도 비대칭 비유사성에 대한 다양한 접근 방법을 소개하고 있으니 참고 바란다.

[프로그램 2.3]과 [결과 2.3]은 〈표 2.7〉 아라비아 숫자에 대한 모스부호의 비대칭 유사성자료에 대한 계량형 다차원척도법을 적용한 것이다. 즉, 〈표 2.7〉은 비대칭 유사성행렬 $C=(c_{ij})$이며, 이로부터 식 (2.12) ~ 식 (2.15)와 같은 개념을 적용할 수 있다. 먼저 비유사성행렬을 대칭과 반대칭행렬에 의해 행렬분해 $C = C_M + C_N = (c_{ij}^M) + (c_{ij}^N)$를 한다. 여기서 $c_{ij}^M = (c_{ij} + c_{ji})/2$와 $c_{ij}^N = (c_{ij} - c_{ji})/2$이다. 다음으로 식 (2.13)과 같이 원소제곱합의 분해 $SSC = SSC_M + SSC_N$를 고려한 후 식 (2.14)와 같은 전체 제곱합 SSC에서 차지하는 SSC_M와 SSC_N의 상대적 비율 $RSSC_M$와 $RSSC_N$을 얻게 된다. 상대적 비율을 통하여 대칭성을 만족하는 C_M의 다차원척도법의 유용성을 살펴보게 된다.

실제로 [결과 2.3]-① Symmetric and Skew-symmetric Similarity matrix from Asymmetric Similarity matrix에서 행렬분해 $C = C_M + C_N$의 대칭과 반대칭행렬을 보여 주고 있다. ② Decomposition of sum of squares due to Symmetry and Skew-symmetry에서는 $RSSC_M$이 99.25%로 $RSSC_N$의 0.75%에 비해 매우 커서 SSC에서 차지하는 SSC_M이의 유용성이 매우 높음을 보여 주고 있다. ③ Dissimilarity matrix from Similarity matrix에서는 대칭성을 만족하는 C_M에 의하여 표준변환하여 얻어진 비유사성행렬을 나타내고 있으며 이에 대한 계량형 다차원척도그림이 ④에 주어져 있다. 이 그림에 대한 해석은 3.3절에서 대칭 유사성행렬 자료 〈표 3.2〉에 대한 계량형 다차원척도그림과 동일하므로 생략한다.

[프로그램 2.3] 모스부호의 비대칭 유사성자료의 계량형 다차원척도법을 위한 프로그램

```
data morsecode;
input code$  x1-x10;
cards;
 1.----        84   63   13    8   10    8   19   32   57   55
 2..---        62   89   53   20    5   14   20   21   16   11
 3...--        16   59   86   31   23   41   16   17   08   10
 4....-         6   23   38   89   42   44   32   10    3    3
 5.....        12    8   27   56   90   42   24   10    6    5
 6-....        12   14   33   34   30   86   69   14    5   14
 7--...        20   25   17   24   18   65   85   70   20   13
 8---..        37   25   16   13   10   22   65   88   61   26
 9----.        57   28    9    7    5    8   31   58   91   78
 0-----        52   18    9    7    5   18   15   39   79   94
;
proc iml;
   reset nolog;

  use morsecode;
  read  all var{code} into code;
  read all var{x1 x2 x3 x4 x5 x6 x7 x8 x9 x10} into C;
  n=nrow(C);

  J=J(n,n,1);
  CM=(C+C`)/2; CN=(C-C`)/2;
  dc=diag(C);
  SSC=C[##]-dc[##]; SSCM=CM[##]-dc[##]; SSCN=CN[##];
  RSSCM=100*SSCM/SSC; RSSCN = 100*SSCN/SSC;
  print, " Symmetric and Skew-symmetric Similarity matrix from
  Asymmetric Similarity matrix",
              Cm[rowname=code colname=code format=5.3]
       Cn[rowname=code colname=code format=5.3];
  print, " Decomposition of sum of squares due to Symmetry and
  Skew-symmetry";
  print, "SSC = SSCM + SSCN", SSC SSCM SSCN RSSCM RSSCN;

  /* Standard Transformation from C to D */
  C=Cm;
  D=sqrt(diag(C)*J -2*C+J*diag(C)); /* d_ij = sqrt[c_ii -2c_ij + c_jj] */
 print, "Dissimilarity matrix from Similarity matrix",
               D[rowname=code colname=code format=5.3];
      /* Creating SAS Data Set for PROC MDS */
    create dism from D[rowname=code];
  append from D[rowname=code];
  close dism;
  quit;

ods graphics on;
/* Metric MDS */
proc mds data=dism level=absolute pfinal;
      id code;
```

[결과 2.2] [프로그램 2.3]의 수행결과

Symmetric and Skew-symmetric Similarity matrix from Asymmetric Similarity matrix

①

CM										
	1.----	2..---	3...--	4....-	5.....	6-....	7--...	8---..	9----.	0-----
1.----	84.00	62.50	14.50	7.000	11.00	10.00	19.50	34.50	57.00	53.50
2..---	62.50	89.00	56.00	21.50	6.500	14.00	22.50	23.00	22.00	14.50
3...--	14.50	56.00	86.00	34.50	25.00	37.00	16.50	16.50	8.500	9.500
4....-	7.000	21.50	34.50	89.00	49.00	39.00	28.00	11.50	5.000	5.000
5.....	11.00	6.500	25.00	49.00	90.00	36.00	21.00	10.00	5.500	5.000
6-....	10.00	14.00	37.00	39.00	36.00	86.00	67.00	18.00	6.500	16.00
7--...	19.50	22.50	16.50	28.00	21.00	67.00	85.00	67.50	25.50	14.00
8---..	34.50	23.00	16.50	11.50	10.00	18.00	67.50	88.00	59.50	32.50
9----.	57.00	22.00	8.500	5.000	5.500	6.500	25.50	59.50	91.00	78.50
0-----	53.50	14.50	9.500	5.000	5.000	16.00	14.00	32.50	78.50	94.00

CM										
	1.----	2..---	3...--	4....-	5.....	6-....	7--...	8---..	9----.	0-----
1.----	0.000	0.500	-1.50	1.000	-1.00	-2.00	-.500	-2.50	0.000	1.500
2..---	-.500	0.000	-3.00	-1.50	-1.50	0.000	-2.50	-2.00	-6.00	-3.50
3...--	1.500	3.000	0.000	-3.50	-2.00	4.000	-.500	0.500	-.500	0.500
4....-	-1.00	1.500	3.500	0.000	-7.00	5.000	4.000	-1.50	-2.00	-2.00
5.....	1.000	1.500	2.000	7.000	0.000	6.000	3.000	0.000	0.500	0.000
6-....	2.000	0.000	-4.00	-5.00	-6.00	0.000	2.000	-4.00	-1.50	-2.00
7--...	0.500	2.500	0.500	-4.00	-3.00	-2.00	0.000	2.500	-5.50	-1.00
8---..	2.500	2.000	-.500	1.500	0.000	4.000	-2.50	0.000	1.500	-6.50
9----.	0.000	6.000	0.500	2.000	-.500	1.500	5.500	-1.50	0.000	-.500
0-----	-1.50	3.500	-.500	2.000	0.000	2.000	1.000	6.500	0.500	0.000

Decomposition of sum of squares Due to Symmetey an Skew-symmetry

②

SSC = SSCM + SSCN

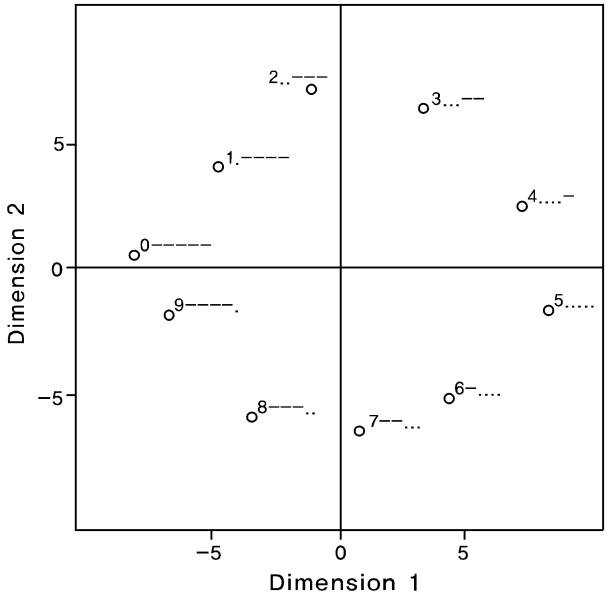
SSC	SSCM	SSCN	RSSCM	RSSCN
99910	99157	753	99.246322	0.7536783

Dissimilarity matrix from Similarity matrix

③

CM										
	1.----	**2..---**	**3...--**	**4....-**	**5.....**	**6-....**	**7--...**	**8---..**	**9----.**	**0-----**
1.----	0.000	6.928	11.87	12.61	12.33	12.25	11.40	10.15	7.810	8.426
2..---	6.928	0.000	7.937	11.62	12.88	12.12	11.36	11.45	11.66	12.41
3...--	11.87	7.937	0.000	10.30	11.22	9.899	11.75	11.87	12.65	12.69
4....-	12.61	11.62	10.30	0.000	9.000	9.849	10.86	12.41	13.04	13.15
5.....	12.33	12.88	11.22	9.000	0.000	10.20	11.53	12.57	13.04	13.19
6-....	12.25	12.12	9.899	9.849	10.20	0.000	6.083	11.75	12.81	12.17
7--...	11.40	11.36	11.75	10.86	11.53	6.083	0.000	6.164	11.18	12.29
8---..	10.15	11.45	11.87	12.41	12.57	11.75	6.164	0.000	7.746	10.82
9----.	7.810	11.66	12.65	13.04	13.04	12.81	11.18	7.746	0.000	5.292
0-----	8.426	12.41	12.69	13.15	13.19	12.17	12.29	10.82	5.292	0.000

④



계량형 다차원척도그림

연습문제

2.1 다음은 5종(species) 분류로 사자, 기린, 소, 양, 인간에 대하여 1=꼬리가 길다, 2=거친 동물이다, 3=긴 목을 가졌다, 4=농가동물이다, 5=다른 동물을 먹이로 한다, 6=네 다리로 걷는다는 6가지 특징으로 분류한 이진수자료이다(Chatfield와 Collins, 1980, p. 194).

종	특징					
	1	2	3	4	5	6
사자	1	1	0	0	1	1
기린	1	1	1	0	0	1
소	1	0	0	1	0	1
양	1	0	0	1	0	1
인간	0	0	0	0	1	0

1) 러셀-라오계수와 단순매칭계수를 개체 간의 짝을 통하여 구하라.
2) 식 (2.5)와 식 (2.6)의 유사성행렬 C_R과 C_S에 관한 행렬식을 이용하여 계산해 보라.
2) 유사성행렬자료 C_S에 대하여 다차원척도법을 하라.

2.2 다음은 Johnson과 Wichern(2002, p. 675)의 5명의 학생에 대한 특성을 조사한 자료를 변형한 것이다.

1) 이 자료를 주관적 기준에 따라 이진수자료를 만들어 보자.
2) 이진수자료에 대해 단순매칭계수를 유사성으로 하는 행렬자료를 제공하라.
3) 이 행렬자료에 대한 다차원척도법을 실시하고 학생들의 군집을 해석하라.
4) 키와 몸무게에 대한 유사성을 곱적률 상관계수와 카이제곱 검정을 통하여 각각 살펴보라.

학생	키(cm)	몸무게(kg)	눈색깔	머리색깔	성별
1	172	65	검정색	갈색	여
2	185	70	갈색	갈색	남
3	170	62	푸른색	검정색	남
4	162	60	갈색	갈색	여
5	193	80	갈색	갈색	남

2.3 1장의 [연습문제 1.4]에서 켈로그회사에 의해서 제조된 총 23종류의 시리얼에 대한 자료로부터 주관적으로 만든 이진수자료에 대하여 다음 질문에 답하라.

1) 유사성 계수로 단순매칭계수를 이용하여 비유사성행렬을 구하고 다차원척도법을 실시하라.
2) 자카드계수나 러셀-라오계수를 이용한 다차원척도법과 1)에서 구한 결과에서 시리얼 유형의 군집을 서로 비교하라.

2.4 〈표 2.4〉로부터 정의한 식 (2.7)의 두 변수의 곱적률 상관계수

$$r = \frac{n_{11}n_{22} - n_{12}n_{21}}{[(n_{11}+n_{12})(n_{21}+n_{22})(n_{11+}n_{21})(n_{12}+n_{22})]^{1/2}}$$

의 제곱이 카이제곱검정 통계량과 수리적으로 연관성이 있음을 보여라.

2.5 다음은 세 가지 재질 돌, 청동, 철을 가지고 만든 5가지 절단공구 T1, T2, T3, T4, T5를 분류한 이진수자료이다. 만약에 한 공구를 만든 특정한 재질이 있다면 1, 없다면 0으로 만든 이진수자료이다(Seber, 1984, p. 252).

절단 공구	재질		
	돌	청동	철
T1	0	0	1
T2	0	1	1
T3	0	1	0
T4	1	1	0
T5	1	0	0

1) 자카드계수에 의한 유사성행렬 C_J가 다음과 같이 계산됨을 확인하라.

T1	1				
T2	0.5	1			
T3	0	0.5	1		
T4	0	0.5	0.5	1	
T5	0	0	0	0.5	1

2) 다차원척도법을 실시하여 유사한 도구를 군집화하라.

2.6 다음은 8가지 직업군(1. 전문직, 2. 자동화 대기업, 3. 자동화 소기업, 4. 자동화 자영업, 5. 수동화 대기업, 6. 수동화 소기업, 7. 수동화 자영업, 8.농업)에 대한 세대 간 직종의 유동성에 대한 비대칭 비유사성자료이다(Cox와 Cox, 2001, pp. 119-121). 이는 i번째 직업을 가진 아버지들에 대한 j번째 직업을 가지는 아들들에 대한 백분율을 x_{ij}라고 하면 비유사성 $d_{ij} = |x_{ii} - x_{ij}|$로부터 얻어진 것이다.

1	0.0	25.3	19.5	40.1	40.5	38.1	43.1	43.8
2	10.9	0.0	9.5	21.7	18.1	22.8	26.4	29.0
3	26.5	17.4	0.0	28.1	26.3	13.9	30.1	34.6
4	19.5	21.9	13.3	0.0	24.3	25.4	26.7	32.1
5	13.2	17.6	16.1	29.9	0.0	20.0	26.5	31.9
6	20.6	15.4	7.5	24.0	10.9	0.0	19.6	26.1
7	25.5	20.1	19.4	23.5	21.3	12.7	0.0	30.3
8	1.7	2.3	0.9	3.4	3.9	10.7	2.6	0.0

1) 식 (2.11)에 의해 비대칭 비유사성행렬 $D = (d_{ij})$의 대칭행렬 D_M과 반대칭행렬 D_N을 구하라.
2) 식 (2.15)에 의해서 전체 제곱합 SSD에서 차지하는 SSD_M와 SSD_N의 상대적 비율 $RSSD_M$이 $RSSD_N$을 구하고 대칭행렬 D_M에 대한 다차원척도법이 타당한지를 논하라.
3) 대칭행렬 D_M에 대한 계량형과 비계량형 다차원척도법의 결과를 비교하라.
4) 반대칭행렬 D_N에 대하여 $D_N^t D_N$의 고유값과 고유벡터를 통한 2차원상의 그림을 얻고 해석하라.

3장
이원 다차원척도법

이원 다차원척도법(two-way MDS)이란 이원행렬 1개의 비유사성행렬 D에 적용된 계량형과 비계량형 다차원척도법을 말한다. 이 장에서는 이들을 위한 알고리즘 중 계량형 다차원척도법을 위한 토거선(Torgerson) 알고리즘과 비계량형 다차원척도법을 위한 크루스칼-세퍼드(Kruskal-Shepard) 알고리즘을 소개하고, 예제를 통해 이를 이해하고 응용을 위해 프로그램을 제공하고 있다. 끝으로 계량형과 비계량형 다차원척도법의 비교를 통하여 그 차이점을 이해하려 한다.

3.1 계량형 다차원척도법

1장에서 살펴보았듯이 Kruskal과 Wish(1978, pp. 22-23)는 비유사성을 나타내는 거리 d_{ij}와 차원축소된 k차원 형상공간의 거리 δ_{ij}의 관계가 절대척도모형, 구간척도모형, 비율척도모형을 만족하도록 하는 다차원척도법을 계량형 다차원척도법이라 하였다. 그리고 순서척도모형의 경우를 비계량형 다차원척도법이라고 한다.

먼저 계량형 다차원척도법을 위한 대표적인 알고리즘은 Young과 Householder(1938)를 바탕으로 기초한 토거선 알고리즘(Torgerson, 1958)이 있다. Gower(1966)는 이를 주성분좌표분석(principal coordinate analysis)이라 하여 더 명료하게 논의하였고 주성분분석과 매우 관련성이 있음을 보였다. 더군다나 이 알고리즘에 의해 얻어지는 저차원 형상공간의 좌표를 다차원척도법에서는 고전적 또는 계량형 해(solution)라고 한다(Seber, 1984, pp. 235-240).

3.1.1 토거선 알고리즘

일반적으로 개체 간의 유클리드거리를 비유사성으로 하는 크기가 $n \times n$인 비유사성행렬 $D = (d_{ij}),\ i, j = 1, \ldots, n$가 있다면 토거선 알고리즘은 [알고리즘 2]와 같이 정리된다(Mardia 외 2인, 1979, p. 400; du Toit 외 2인, 1986, p. 128; Davison, 1992, pp. 63-64).

특히, Davison(1992, Chapter 4)은 [알고리즘 2: 토거선 알고리즘]이 유클리드공간에서 개체 간의 유클리드거리를 비유사성으로 가정하고 있어 실제 다양한 거리에 바탕을 두는 응용에서는 변형된 알고리즘이 사용되고 있음을 지적하고 있다. 실제로 Mardia 외 2인

[알고리즘 2] 토거선 알고리즘
1단계: 비유사성행렬 $D = (d_{ij}),\ i, j = 1, \ldots, n$ 로부터 행렬 $A = (a_{ij})$를 계산한다. 여기서 $a_{ij} = -d_{ij}^2/2$.
2단계: 행렬 A 로부터 이중-중심화행렬 $B = (b_{ij}) = HAH$를 계산한다. 여기서, $b_{ij} = a_{ij} - \overline{a_{i.}} - \overline{a_{.j}} + \overline{a_{..}}$ 이며, $\overline{a_{i.}}, \overline{a_{.j}}$ 그리고 $\overline{a_{..}}$는 각각 행렬 A의 i 번째 행의 평균, j 번째 열의 평균 그리고, 모든 원소의 평균이다. $H = I_n - J_n/n$이며 I_n은 단위행렬이고 J_n은 모든 원소가 1인 $n \times n$ 행렬이다.

3단계: 행렬 B의 스펙트럼분해 $B = VD_{\lambda}V^{t}$를 계산한다. 여기서, D_{λ}는 고유값 $\lambda_1 \geqq \cdots \geqq \lambda_p > 0$를 대각원소로 하는 대각행렬이며, V는 고유벡터 $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$가 열인 직교행렬이다.

4단계: 행렬 B의 스펙트럼분해로부터 $k\,(\leqq p)$개의 고유값과 이에 대응하는 고유벡터를 가지고 크기가 $n \times k$인 행렬

$$X_{(k)} = V_{(k)} D_{\lambda(k)}^{1/2} = (\boldsymbol{v}_1 \lambda_1^{1/2}, \cdots, \boldsymbol{v}_k \lambda_k^{1/2})$$

를 계산한다. 이는 k차원의 형상공간의 좌표를 제공한다.

(1979, Section 14.4)은 비유사성이 유클리드거리가 아니더라도 근사적으로 [알고리즘 2]가 적용될 수 있음을 보여 주고 있다. 사실 1.3.1절에서 설명한 측정수준에 근거한 절대척도모형과 5장의 Ramsay(1977, 1980, 1982)의 최대우도 다차원척도법도 여기에 해당된다.

3.1.2 형상공간의 적합도

마지막 [4단계]에서 제공되는 k차원의 형상공간의 좌표가 k차원의 다차원척도그림을 제공한다. 만약 λ_1이 $\lambda_2, \ldots, \lambda_p$보다 훨씬 크다면 1차원의 형상공간에서 개체들의 거리가 비유사성행렬 D의 개체들의 비유사성인 거리를 근사적으로 나타낸다. 따라서 적절한 차원의 수 k를 결정하는 것도 중요한 과제이며 이와 관련하여 이미 1.3.3절에서 스트레스-차원수 그림과 차원수 결정을 위한 크루스칼의 판별기준 〈표 1.9〉를 소개한 바 있다. 여기서는 Mardia 외 2인(1979, pp. 407-408)이 [알고리즘 2]를 토대로 k차원 형상공간의 $X_{(k)}$에 의해서 설명되는 비유사성행렬 D의 설명력인 두 가지 적합도(goodness of fit)를

$$\alpha_{1,k} = \left(\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{n} |\lambda_i|\right) \times 100\%, \quad \lambda_k > 0, \tag{3.1}$$

$$\alpha_{2,k} = \left(\sum_{i=1}^{k} \lambda_i^2 \Big/ \sum_{i=1}^{n} \lambda_i^2\right) \times 100\%, \quad \lambda_k > 0 \tag{3.2}$$

제안하고 있다. 식 (3.1)에서 절댓값을 사용한 것은 이들 고유값 중 일부가 작은 것은 음이 되는 경우를 고려한 것이다. 특히 $\alpha_{1,k}$는 주성분분석에서 고유값의 전체합계에 대한 최대 고유값부터 내림순으로 몇 개의 고유값이 차지하는 비율을 보고서 고려할 만한 고유값에 대응되는 주성분의 수를 정하는 기준과 대동소이함을 알 수 있다. 그리고 적절한 차원의 수 k를 정하는 기준으로 Davison(1992, pp. 68-70)은 수직축은 고유값을 수평축에는 차원수를 나타낸 그림을 권하고 있다. 이는 일반적으로 인자분석에서 쉽게 찾아볼 수 있다.

Mardia 외 2인(1979, p. 405)은 실제 자료분석 상황에서는 [알고리즘 2]의 [1단계]에서 크기가 $n \times n$ 비유사성행렬 D로 시작하는 대신에 크기가 $n \times p$인 다변량자료행렬 $X = (x_{ir})$, $i = 1, \ldots, n; r = 1, \ldots, p$로 시작하게 됨을 강조한다. 만약에 $\lambda_1 \geqq \cdots \geqq \lambda_p > 0$를 크기가 $p \times p$인 표본 공분산행렬 S로부터 얻어지는 $nS = X^t HX$의 고유값이라고 하자. 이는 [2단계] $B = HXX^t H$의 고유값이 되며 $B = (b_{ij})$는 중심화 내적행렬로

$$b_{ij} = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^t (\boldsymbol{x}_j - \overline{\boldsymbol{x}})$$

이다. [3단계]인 행렬 B의 스펙트럼분해로부터 $k\,(\leqq p)$개의 고유값과 이에 대응하는 고유벡터로 이루어진 $V_{(k)} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k)$의 행들은 직교행렬로 다변량자료행렬 X의 $k\,(\leqq p)$ 차원의 주성분좌표가 된다.

토거선 알고리즘의 이론적이며 수리적 증명은 Mardia 외 2인(1979, Chapter 4)와 Davison(1992, pp. 73-76)을 참고하기를 바란다.

3.1.3 계량형 다차원척도법의 예

이 절에서는 이미 비계량형 다차원척도법이 적용된 1.1.2절 [사례 2] 우리나라 14개 경제관련기관의 경제 전망에 대하여 3.1.1절의 [알고리즘 2]의 토거선 알고리즘을 바탕으로 하는 계량형 다차원척도법을 적용하기 위한 SAS/IML 프로그램을 주로 소개하려고 한다. 이로부터 얻어진 형상좌표를 통하여 SAS/PROC GPLOT과 결합하여 2차원 다차원척도그림을 얻게 되며 이를 위해 [프로그램 3.1]과 [결과 3.1]을 제공하고 있다.

[프로그램 3.1] [알고리즘 2]를 위한 SAS/IML과 PROC GPLOT 프로그램

```
/* Metric MDS based on Torgerson's algorithm */
PROC IML;
   RESET NOLOG;

/* SAS/IML Module for MMDS */

/* 1) Double-Centered Matrix: B */
START MMDS;
   n = nrow(D);
   A = -D#D/2;
   J = J(n,n,1);
   H = I(n)-J/n;
   B = H*A*H;
   print, " Double-Centered Matrix", B ;

/* 2) Spectral Decomposition of B */
   call eigen(e, V, B);
         absle = abs(e);
         sum = sum(absle);
         percent = absle / sum * 100;
         G_of_fit = (percent[1,1] + percent[2,1]);
print, "eigenvalue  proportion  Goodness of fit", absle percent G_of_fit;
/* 3) Making a vector with positive eigenvalues */
   s=0;
   do k= 1 to n;
      if e[k] > 0 then s=s+1;
   end;

abse=J(s,2);

   do k=1 to s;
      abse[k,1]=k;
      abse[k,2]=absle[k];
   end;

/* 4) Configuration */
   X = V*diag(sqrt(absle));
   X2 = X[,1:2];
   print , "Configuration", X2;

FINISH;
/* 5) Dissimilarity  Matrix */
```

```
D={.0 .4 .4 .4 .5 .5 .8 .3 .6 .4 .6 .3 .8 .7,
   .4 .0 .2 .8 .5 .3 .4 .3 .2 .8 .2 .3 .6 .5,
   .4 .2 .0 .8 .5 .5 .6 .5 .2 .8 .4 .1 .8 .7,
   .4 .8 .8 .0 .5 .7 .6 .5 .8 .0 .6 .7 .4 .5,
   .5 .5 .5 .5 .0 .4 .5 .4 .5 .5 .7 .4 .5 .4,
   .5 .3 .5 .7 .4 .0 .3 .6 .5 .7 .5 .4 .5 .2,
   .8 .4 .6 .6 .5 .3 .0 .5 .6 .6 .4 .7 .2 .1,
   .3 .3 .5 .5 .4 .6 .5 .0 .5 .5 .5 .6 .5 .6,
   .6 .2 .2 .8 .5 .5 .6 .5 .0 .8 .2 .3 .6 .7,
   .4 .8 .8 .0 .5 .7 .6 .5 .8 .0 .6 .7 .4 .5,
   .6 .2 .4 .6 .7 .5 .4 .5 .2 .6 .0 .5 .4 .5,
   .3 .3 .1 .7 .4 .4 .7 .6 .3 .7 .5 .0 .9 .6,
   .8 .6 .8 .4 .5 .5 .2 .5 .6 .4 .4 .9 .0 .3,
   .7 .5 .7 .5 .4 .2 .1 .6 .7 .5 .5 .6 .3 .0};
RUN MMDS;

 /* 6) Convert the matrix of IML into the external data set */
   vname={dim1 dim2};
   create config from X2 [colname=vname];
   append from X2;
   close config;
   QUIT;

/* 7) Labels & configuration for MMDS plot */
data id;
      input id $ @@;
      cards;
hankook daewoo1 dongseo1 kyungrun sankong choongso hyundai
samsung sunkyung daewoo2 shinhan dongseo2 korea chungang
;
data config;
      merge id config;
proc print data=config;

/* 8) 2-dimensional MMDS plot */
goptions reset=all;
symbol value=circle pointlabel=("#id");
axis1 order=(-0.5 to 0.5 by 0.5) length=5 in;
proc gplot data=config;
      plot dim2 * dim1/ haxis=axis1 vaxis=axis1 href=0 vref=0 frame;
```

[프로그램 3.1]은 8개 과정으로 구성되어 있는데, 새로운 비유사성행렬을 분석하고자 할 때는 단지 /* 5) Dissimilarity Matrix */의 D={....}와 /* 7) Labels & configuration for MMDS plot */의 data 문의 개체를 나타내는 변수 id의 라벨만 수정하면 된다. 경우에 따라 〈proc print data=config;〉의 실행 결과인 좌표값을 보고 /* 8) 2-dimensional MMDS plot */의 axis1 order=(-0.5 to 0.5 by 0.5)에서 범위를 새롭게 지정해야 할 것이다. 여기에서는 [알고리즘 2]를 적용하기 위한 사례로 1.1.2절의 [사례 2]의 우리나라 14개 경제관련기관의 비유사성행렬 〈표 1.5〉를 사용하고 있다.

/* 1) Double-Centered Matrix: B */ 는 [알고리즘 2]의 [1단계]와 [2단계]를 나타내고 실행결과는 [결과 3.1]-①에 주어져 있다.

[결과 3.1] [프로그램 3.1]의 수행결과

①

Double-Centered Matrix

B

	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9	COL10	COL11	COL12	COL13	COL14
ROW1	0.1410714	0.0267857	0.0653571	0.0925	-0.008214	-0.008214	-0.188929	0.0732143	-0.038929	0.0925	-0.063214	0.0975	-0.166071	-0.115357
ROW2	0.0267857	0.0725	0.0910714	-0.181786	-0.0425	0.0375	0.0167857	0.0389286	0.0867857	-0.181786	0.0625	0.0632143	-0.060357	-0.029643
ROW3	0.0653571	0.0910714	0.1496429	-0.143214	-0.003929	-0.003929	-0.044643	-0.0025	0.1253571	-0.143214	0.0410714	0.1417857	-0.161786	-0.111071
ROW4	0.0925	-0.181786	-0.143214	0.2039286	0.0232143	-0.096786	-0.0175	0.0246429	-0.1475	0.2039286	-0.031786	-0.071071	0.1053571	0.0360714
ROW5	-0.008214	-0.0425	-0.003929	0.0232143	0.0925	0.0125	-0.018214	0.0139286	-0.008214	0.0232143	-0.1525	0.0382143	0.0046429	0.0253571
ROW6	-0.008214	0.0375	-0.003929	-0.096786	0.0125	0.0925	0.0617857	-0.086071	-0.008214	-0.096786	-0.0325	0.0382143	0.0046429	0.0853571
ROW7	-0.188929	0.0167857	-0.044643	-0.0175	-0.018214	0.0617857	0.1210714	-0.016786	-0.048929	-0.0175	0.0267857	-0.1125	0.1239286	0.1146429
ROW8	0.0732143	0.0389286	-0.0025	0.0246429	0.0139286	-0.086071	-0.016786	0.0953571	-0.006786	0.0246429	-0.031071	-0.060357	0.0060714	-0.073214
ROW9	-0.038929	0.0867857	0.1253571	-0.1475	-0.008214	-0.008214	-0.048929	-0.006786	0.1410714	-0.1475	0.0967857	0.0975	-0.026071	-0.115357
ROW10	0.0925	-0.181786	-0.143214	0.2039286	0.0232143	-0.096786	-0.0175	0.0246429	-0.1475	0.2039286	-0.031786	-0.071071	0.1053571	0.0360714
ROW11	-0.063214	0.0625	0.0410714	-0.031786	-0.1525	-0.0325	0.0267857	-0.031071	0.0967857	-0.031786	0.0925	-0.006786	0.0496429	-0.019643
ROW12	0.0975	0.0632143	0.1417857	-0.071071	0.0382143	0.0382143	-0.1125	-0.060357	0.0975	-0.071071	-0.006786	0.1439286	-0.249643	-0.048929
ROW13	-0.166071	-0.060357	-0.161786	0.1053571	0.0046429	0.0046429	0.1239286	0.0060714	-0.026071	0.1053571	0.0496429	-0.249643	0.1667857	0.0975
ROW14	-0.115357	-0.029643	-0.111071	0.0360714	0.0253571	0.0853571	0.1146429	-0.073214	-0.115357	0.0360714	-0.019643	-0.048929	0.0975	0.1182143

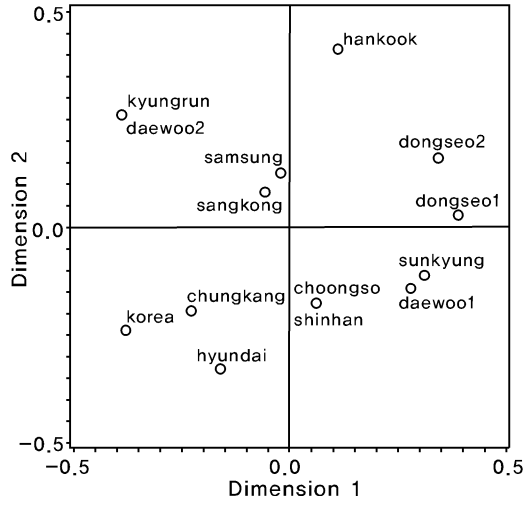
②

eigenvalue proportion Goodness of fit

absle	percent	G_of_fit
0.9918302	34.205809	56.666533
0.6512702	22.460724	
0.3357433	11.578969	
0.1902143	6.5600275	
0.140498	4.8454324	
0.0520351	1.7945622	
0.0057069	0.1968161	
4.605E-17	1.588E-15	
3.53E-17	1.218E-15	
0.0414413	1.4292091	
0.0712893	2.458596	
0.0996532	3.4367968	
0.1316732	4.5410882	
0.1882409	6.4919699	

OBS	id	DIM1	DIM2
1	hankook	0.11501	0.41472
2	daewoo1	0.28487	-0.14126
3	dongseo1	0.38989	0.03242
4	kyungrun	-0.38263	0.26388
5	sangkong	-0.04057	0.08268
6	choongso	0.06736	-0.16944
7	hyundai	-0.16227	-0.32930
8	samsung	-0.01688	0.12132
9	sunkyung	0.31341	-0.11213
10	daewoo2	-0.38263	0.26388
11	shinhan	0.07181	-0.16150
12	dongseo2	0.34812	0.16601
13	korea	-0.38007	-0.23893
14	chungang	-0.22541	-0.19234

③



/* 2) Spectral Decomposition of B */는 [알고리즘 2]의 [3단계]를 나타내며 3.1.2절에서 정의한 각 고유값과 그들이 전체에서 차지하는 비율 외에 적절한 차원수 k의 결정을 위한 기준이 되는 식 (3.1)의 적합도 $\alpha_{1,k}$도 제공한다. 이 과정의 실행결과는 [결과 3.1]-② 이고 결과를 보면 적합도(G-of-Fit) 56.67%는 2차원 다차원척도법이 비유사성행렬 D를 설명하는 정도를 나타낸다. 매우 높은 적합도는 아니지만 2차원 다차원척도그림의 해석에는 큰 무리가 없다. /* 3) Making a vector with positive eigenvalues */는 앞의 프로그램 과정 2)의 행렬 B의 스팩트럼분해에서 일부 고유값이 0과 음이 되는 경우가 발생하므로 이들을 제외한 양의 고유값을 갖는 벡터로 만들어 주는 과정이다.

/* 4) Configuration */은 [알고리즘 2] 마지막 [4단계]에서 $k=2$ 차원의 형상좌표를 제공하며 [결과 3.1]-②의 오른쪽에 주어져 있다. /* 6) Convert the matrix of IML into the external data set */은 프로그램 과정 3)과 과정 4)에서 만들어진 SAS/IML 내의 형상좌표를 /* 7) Labels & configuration for MMDS plot */에서 사용하기 위한 SAS 데이터 세트 〈config〉를 만들어 준다.

마지막 단계로 [결과 3.1]-③의 2차원 다차원척도그림은 프로그램의 /* 8) 2-dimensional MMDS plot */의 과정으로부터 얻을 수 있다. 이는 1.1.2절의 〈그림 1.2〉의 (b) 이진수자료의 제곱유클리드거리에 의한 비계량형 다차원척도그림과 해석이 동일하며 개체 간의 군집도 대동소이함을 알 수 있다. 단지 두 그림에서 형상좌표의 척도에 차이점이 있지만 다차원척도그림의 해석에는 아무런 문제가 되지 않는다.

3.2 비계량형 다차원척도법

이 절에서는 비계량형 다차원척도법을 위한 대표적인 크루스칼-세퍼드(Kruskal-Shepard)의 알고리즘을 소개하려고 한다. 먼저 1.3.1절에서 비계량형 다차원척도법에 관련된 순서척도모형 $d_{ij}=f(\delta_{ij})+\epsilon_{ij}$를 다시 생각해 보자. 모형에서 함수 $f(\cdot)$는

$$\text{모든 } i,j,k,l=1,\ldots,n\text{에 대하여 } d_{ij}\le d_{kl}\text{이면 } f(d_{ij})\le f(d_{kl})\text{가 성립}$$

하는 단조증가함수이므로 차원축소된 형상공간의 개체 간의 거리 δ_{ij}를 구하기 위해 사용할 수 있는 정보로 유클리드공간의 개체 간의 거리 d_{ij}들의 크기순서(rank order)가 유일하다. 예를 들어 1.1.1절의 [사례 1]에서 우리나라 도시 간 철도거리로 지도를 만든다고 했을 때

다음과 같은 정보를 활용하게 된다.

인천에서 서울까지의 거리가 30.9km로 가장 가깝다.
부산에서 대구까지의 거리가 120.6km로 그 다음 가장 가깝다.
…
광주에서 강릉까지의 거리가 523km로 가장 멀다.

따라서 Seber(1984, p. 241)는 비계량형 다차원척도법의 목표는 d_{ij}와 δ_{ij}에 대한 그림이 (근사적) 단조증가하는 k차원의 형상공간을 찾는 것이라고 하였다. 예를 들어, 만약에 $n=3$인 경우 $d_{23} < d_{12} < d_{13}$라면 $\delta_{23} \leq \delta_{12} \leq \delta_{13}$를 만족하는 3개체들의 형상좌표를 구하는 것이다. 이를 위한 대표적인 알고리즘으로 3.2.1절의 크루스칼-세퍼드 알고리즘이 있다.

3.2.1 크루스칼-세퍼드 알고리즘

비계량형 다차원척도법에서는 비유사성행렬 D가 반드시 거리행렬이라 여기지 않아도 되며 실제로 자료가 유사성행렬로 나타날 때 가장 적절한 경우가 많다. 이와 같은 전제에서 출발하여 1960년대 초에 Shepard(1962)와 Kruskal(1964)에 의해서 비계량 다차원척도법의 형상공간을 제공하는 크루스칼-세퍼드 알고리즘이 개발되었다. 여기서는 Mardia 외 2인(1979, p. 414)과 du Toit 외 2인(1986, pp. 131-133)에 나타나 있는 이 알고리즘을 [알고리즘 3: 크루스칼-세퍼드 알고리즘]으로 소개하기로 한다.

알고리즘의 [4단계]에서 스트레스 공식의 최소화는 1.3.2절의 스트레스에서 설명한 [알고리즘 1: 스트레스 최소화 알고리즘]을 따른다. 그리고 [3단계]의 순위상 계산과정을 최소제곱단조회귀(least square monotone regression)라고 한다. Guttman(1968)이 순위상(rank image)이라는 용어를 사용하였고 불일치(disparity)라고도 한다(Izenman, 2008, pp. 492-493). 그리고 [4단계]의 스트레스는 1.3.2절의 식 (1.7)의 $ST1$과 궁극적으로 같아 이를 크루스칼의 〈스트레스 1〉이라고 한다. 따라서 식 (3.3)의 $\hat{d}_{ij}$은 $ST1$의 $f(\cdot)$의 최적(적합된)값을 의미한다. 스트레스는 이외에도 $ST2$와 Young의 S-스트레스 $SST1$과 $SST2$가 있는데 1.3.2절에 잘 설명하고 있다. 그리고 [알고리즘 3]에서 다양한 스트레스 공식의 적용은 다차원척도법을 위한 다양한 형상좌표를 제공하게 된다.

[알고리즘 3] 크루스칼-세퍼드 알고리즘
1단계: 크기가 $n \times n$인 비유사성행렬 $D=(d_{ij})$로부터 $m=n(n-1)/2$개의 비대각원소를 그 크기에 따라 다음과 같이 순서화한다. $$d_{i_1j_1} \le d_{i_2j_2} \le \ \ldots \ \le d_{i_mj_m}$$ **2단계:** X^*를 k차원 형상공간의 좌표행렬이라 하고 이에 대응하는 비유사성행렬을 D^*라 한다면 $\delta^*_{i_1j_1}, \delta^*_{i_2j_2}, \ldots, \delta^*_{i_mj_m}$는 당연히 $d_{i_1j_1}, d_{i_2j_2}, \ldots, d_{i_mj_m}$에 대응되는 값이다. **3단계:** $\delta^*_{i_1j_1}, \delta^*_{i_2j_2}, \ldots, \delta^*_{i_mj_m}$와 거의 일치하면서 $d_{i_1j_1}, d_{i_2j_2}, \ldots, d_{i_mj_m}$에 단조관계를 갖는 순위상(rank image)인 $\hat{d}_{i_1j_1} \le \hat{d}_{i_2j_2} \le \ldots \le \hat{d}_{i_mj_m}$을 계산한다. **4단계:** 크루스칼의 스트레스 $$\left[\sum_{i<}\sum_{j}(\delta^*_{ij}-\hat{d}_{ij})^2 / \sum_{i<}\sum_{j}(\delta^*_{ij})^2\ \right]^{1/2} \qquad (3.3)$$ 을 정의하고 이를 최소화하는 형상공간 좌표행렬 X^*를 찾는다.

Davison(1992, Chapter 5)에서 크루스칼의 스트레스를 최소화하는 비계량형 다차원척도법을 제공하는 패키지 프로그램(MDSCAL, TORSCA, KYST)의 특징을 소개하고 있다. 이들 중 가장 최근의 KYST가 나머지의 장점을 포함하고 있고 특히 사용자가 스트레스 1과 2를 선택할 수 있다고 하였다. 그리고 ALSCAL(Takane 외 2인, 1977; Young과 Lewyckyj, 1979)은 사용자가 $SST1$과 $SST2$를 선택할 수 있게 되어 있고 이를 범용 SAS/PROC ALSCAL이 채택하였다. 특별히 Schiffman 외 2인(1981)은 KYST와 ALSCAL의 알고리즘과 사용법에 대해 설명하고 있다. SAS/PROC MDS에 모든 스트레스를 사용자가 옵션 〈FIT=〉으로 선택할 수 있고, 이는 [부록 1: SAS MDS 절차]에서 자세히 설명하고 있다.

3.2.2 상그림과 세퍼드그림

[알고리즘 3: 크루스칼-세퍼드 알고리즘]의 [2단계]에서 적절한 차원수 k를 정하는 적합도 문제를 다루기로 하자. 일반적으로 가능한 차원수 k를 작게 잡는 것이 해석상 유리하지만 차원수 k가 커질수록 스트레스는 작아지는 경향이 있음을 1.3.3절 차원의 수에서 설명하였다. 또한 차원수 결정을 위한 방법으로는 먼저 크루스칼의 판별기준과 스트레스-차원수 그림을 1.3.3절에서 이미 살펴보았고 3.1.2절의 [알고리즘 2: 토거선 알고리즘]에 바탕을 두는 계량형 다차원척도법에서는 식 (3.1)과 식 (3.2)의 적합도를 정의하였다. 나머지 방법으로는 Guttman(1968)의 상그림(image diagram)과 세퍼드(Separd)그림이 있다. 상그림은 수직축에 순위상($\widehat{d_{ij}}$)과 수평축에 k차원의 형상좌표로부터 얻어지는 비유사성(δ^*_{ij})에 대한 산점도이다. 이는 비계량형 다차원척도법을 위한 순서척도모형의 적합성을 보여 준다. 따라서 이 상그림의 점들이 원점에서 수평축에 대해 45° 각도로 직선상에 놓여있다면 비계량형 다차원척도법에 관한 순서척도모형이 완벽하게 적합되었음을 의미한다. 그리고 이 직선에서 눈에 띄게 벗어나는 점은 이상치(outlier)이다. 이상치가 많을수록 크루스칼의 $ST1$은 양적으로 커지게 된다. 그 이유는 식 (3.3)을 살펴보면 쉽게 알 수 있다.

다음으로 세퍼드그림은 수직축에 순위상($\widehat{d_{ij}}$)과 수평축에 실제 비유사성(d_{ij})에 대한 산점도이다. 상그림과 같이 이 그림에도 순위상이 단조관계를 보여 주기 때문에 순서척도모형에서 가정한 $f(\cdot)$의 단조성을 평가하는 방법이다. 이 그림들의 보다 자세한 설명은 Davison(1992, pp. 95-99)을 참고 바라고 실제 활용의 예를 다음의 3.2.3절로 넘긴다.

3.2.3 비계량형 다차원척도법의 예

3.2.1절과 3.2.2절에서 비계량형 다차원척도법을 위한 [알고리즘 3: 크루스칼-세퍼드 알고리즘]과 k차원 공간좌표의 적합도 문제를 다루었다. 일반적으로 [알고리즘 3]의 다양한 수치적 계산에 따라 MDSCAL, TORSCA, KYST, ALSCAL 등 몇 가지 프로그램이 있음을 3.2.1절에서 설명한 바 있다. 이 절에서는 특별히 범용 SAS/PROC ALSCAL과 대동소이한 PC-윈도우용 SAS/PROC MDS를 이용하여 비계량형 다차원척도법의 실제 활용의 예를 제공하고자 한다. 이미 1.1.2절 [사례 2] 우리나라 14개 경제관련기관의 경제 전망과 1.1.3절 [사례 3] 12개 국가의 유사성에 대한 평가에서 비계량형 다차원척도법을 적용한 바 있다.

〈표 3.1〉 9가지 직업 간의 비유사성행렬자료

Preacher	0.00								
Surgeon	3.06	0.00							
Teacher	2.14	3.95	0.00						
Journalist	3.21	3.95	3.03	0.00					
Policeman	3.51	4.17	2.82	3.33	0.00				
Plumber	4.40	3.77	3.86	4.14	3.60	0.00			
Farmer	3.64	3.69	3.47	3.90	3.56	2.53	0.00		
Actor	3.12	4.13	2.72	2.58	3.68	4.17	4.10	0.00	
Bank clerk	3.73	4.05	3.31	3.46	3.29	3.92	3.59	3.72	0.00

〈표 3.1〉은 9가지 직업 preacher(목사), surgeon(외과의사), teacher(초등학교 교사), journalist(기자), policeman(경찰관), plumber(배관공), farmer(농부), actor(배우), bank clerk(은행원)간의 비유사성행렬자료이다. 이 자료는 62명의 사람에게 9가지 직업을 각각 두 직업씩 짝지은 36가지의 경우에 대해 서로 얼마나 다른지를 0점(유사하다) ~ 5점(완전히 다르다)의 척도로 응답하게 한 후 그 평균으로 얻어진 것이다(du Toit 외 2인, 1986, p. 128). du Toit 외 2인(1986, Chapter 6)에서는 이 자료에 대해 계량형과 비계량형 다차원척도법을 모두 제공하고 있으며 군집분석의 결과도 제공하고 있다. 여기서 전자는 3.1.1절의 [알고리즘 2: 토거선 알고리즘]을 바탕으로 범용 SAS/PROC MATRIX의 프로그램을 이용하였고, 후자는 [알고리즘 3: 크루스칼-세퍼드 알고리즘]를 위한 범용 SAS/PROC ALSCAL 프로그램을 이용하여 제공되고 있다.

이제는 본격적으로 〈표 3.1〉의 자료의 비계량형 다차원척도법을 위한 SAS/PROC MDS [프로그램 3.2]와 [결과 3.2]를 살펴보자.

[프로그램 3.2] [알고리즘 3]의 비계량형 다차원척도법을 위한 프로그램

```
1)    data occupy;
            input (preach surgeon teacher journal police
                   plumber farmer actor bank) (5.2) @60;
      cards;
       0.00
       3.06    0.00
       2.14    3.95    0.00
       3.21    3.95    3.03    0.00
       3.51    4.17    2.82    3.33    0.00
       4.40    3.77    3.86    4.14    3.60    0.00
       3.64    3.69    3.47    3.90    3.56    2.53    0.00
       3.12    4.13    2.72    2.58    3.68    4.17    4.10    0.00
       3.73    4.05    3.31    3.46    3.29    3.92    3.59    3.72    0.00
```

```
2)   ods graphics on;
     proc mds data=occupy level=ordinal nophist pfinal outres=res;

3)    goptions reset=all;
      symbol value=circle;
      axis1 length=5 in;
      proc gplot data=res;
      plot fitdata * data/frame hzero vzero haxis=axis1 vaxis=axis1;

4)   proc print data=res;
                var _row_ _col_ data fitdata fitdist residual;
      run;
```

[프로그램 3.2]는 편의성을 위해 크게 네 과정으로 구성되어 있다. 과정 1)과 2)는 1.4절에서 이미 보여 준 전형적인 PROC MDS 문이다. 과정 3)은 세퍼드그림을 위한 프로그램이다. 이들에 대한 수행결과는 [결과 3.2]에 주어져 있다. 과정 4)는 [결과 3.3]에 따로 정리해 두어 [알고리즘 3]을 수치적으로 확인하고 이해하려 한다.

먼저 [프로그램 3.2]-2)의 proc mds 문에서는 비계량형 다차원척도법을 위한 옵션 〈level= ordinal〉(디폴트)이 주어져 있다. 〈outres= res〉는 3.2.2절에서 설명한 Guttman의 상그림과 세퍼드그림을 위한 자료를 새로운 SAS 자료세트 res에 저장하라는 옵션이다. 이를 과정 3)과 4)에서 활용하게 된다. 9.2판부터는 상그림은 자동으로 제공되므로 이에 대한 프로그램은 생략해도 무방하다. 다만 세퍼드그림을 위한 과정 3)은 필요하다. 그러나 엄밀히 말해서 이들 두 그림의 기능과 역할이 유사하므로 다차원척도법을 처음 다루는 이용자들에게는 과정 3)의 프로그램을 생략해도 다차원척도법 모형의 적합성을 평가하고 해석하는 데 별 무리가 없다고 여겨진다.

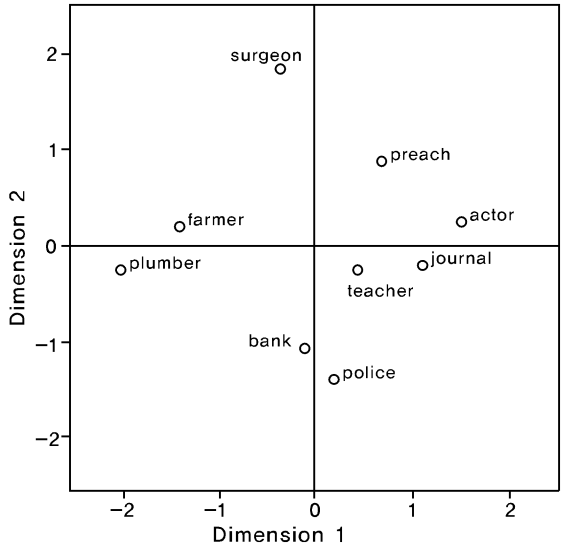
과정 1) ~ 과정 3)의 수행결과는 [결과 3.2]에 요약되어 있다. ①의 (a) 다차원척도그림과 (b) 형상좌표, ②의 〈Badness-of-Fit Criterion〉 그리고 ③의 (a) 상그림과 (b) 세퍼드그림이다. 특히, ②의 〈Badness-of-Fit Criterion〉 값 0.10은 ①의 2차원 다차원척도그림이 개체 간의 관계를 설명하는 정도가 차원수 결정을 위한 크루스칼의 판별기준 〈표 1.9〉에 따르면 〈보통(fair)〉이 됨을 말한다.

이 다차원척도그림에 나타나는 9가지 직업들은 3가지 동질적인 직업군으로 나누어진다. 그림의 수평축(1차원축)의 왼쪽에 육체적 노동력이 요구되는 직업인 farmer(농부)와 plumber(배관공), 오른쪽에는 전문지식이 요구되는 teacher(초등학교 교사), actor(배우), journal(기자)이 각각 그룹을 이루고 있다. 수직축(2차원축)의 아래쪽에는 정복차림의 bank(은행원)와 police(경찰관)가 한 그룹을 이루고 있다. 그 반대편에는 치료와 치유를 직업으로 하

는 surgeon(외과의사)과 preach(목사)가 놓여 있지만 이들의 직업은 서로 비유사할 뿐만 아니라 다른 직업군과도 유사하지 않다.

[결과 3.2] [프로그램 3.2]-1)-3)의 수행결과

①



(a) 다차원척도그림

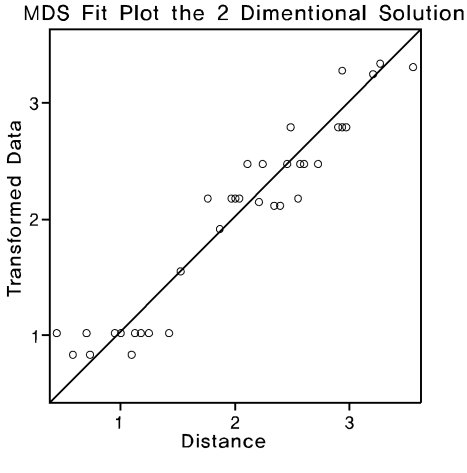
Configuration		
	Dim1	Dim2
preach	0.68	0.85
surgeon	-0.38	1.84
teacher	0.42	-0.24
journal	1.13	-0.19
police	0.21	-1.39
plumber	-2.05	-0.25
farmer	-1.41	0.19
actor	1.50	0.25
bank	-0.09	-1.07

(b) 형상좌표

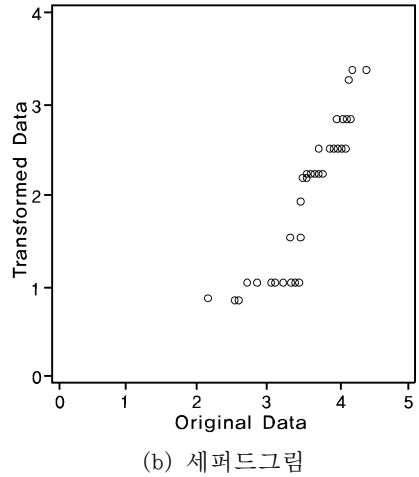
②

MATRIX	Number of Nonmissing Data	Weight	Badness-of-Fit Criterion	Distance Correlation	Uncorrected Distance Correlation
1	36	1.00	0.10	0.96	0.99

③



(a) 상그림



(b) 세퍼드그림

[결과 3.3] [프로그램 3.2]-4)의 수행결과

OBS	_ROW_	_COL_	DATA	FITDATA	FITDIST	RESIDUAL
1	surgeon	preach	3.06	1.02105	1.44773	-0.42668
2	teacher	preach	2.14	0.83875	1.12308	-0.28433
3	teacher	surgeon	3.95	2.46380	2.22910	0.23469
4	journal	preach	3.21	1.02105	1.13515	-0.11411
5	journal	surgeon	3.95	2.46380	2.52873	-0.06494
6	journal	teacher	3.03	1.02105	0.70930	0.31175
7	police	preach	3.51	2.15061	2.28826	-0.13765
8	police	surgeon	4.17	3.30431	3.28270	0.02161
9	police	teacher	2.82	1.02105	1.16605	-0.14500
10	police	journal	3.33	1.50806	1.50635	0.00171
11	plumber	preach	4.40	3.30431	2.94298	0.36133
12	plumber	surgeon	3.77	2.46380	2.67733	-0.21354
13	plumber	teacher	3.86	2.46380	2.46881	-0.00501
14	plumber	journal	4.14	3.21273	3.17692	0.03581
15	plumber	police	3.60	2.17187	2.53187	-0.36000
16	farmer	preach	3.64	2.17187	2.19501	-0.02314
17	farmer	surgeon	3.69	2.17187	1.95297	0.21890
18	farmer	teacher	3.47	1.90292	1.87961	0.02331
19	farmer	journal	3.90	2.46380	2.56616	-0.10237
20	farmer	police	3.56	2.15061	2.26040	-0.10979
21	farmer	plumber	2.53	0.83875	0.77173	0.06702
22	actor	preach	3.12	1.02105	1.01941	0.00163
23	actor	surgeon	4.13	2.79613	2.46271	0.33342
24	actor	teacher	2.72	1.02105	1.18982	-0.16877
25	actor	journal	2.58	0.83875	0.58045	0.25830
26	actor	police	3.68	2.17187	2.08623	0.08564
27	actor	plumber	4.17	3.30431	3.58747	-0.28316
28	actor	farmer	4.10	2.79613	2.91522	-0.11909
29	bank	preach	3.73	2.17187	2.07225	0.09961
30	bank	surgeon	4.05	2.79613	2.92704	-0.13091
31	bank	teacher	3.31	1.02105	0.97533	0.04572
32	bank	journal	3.46	1.50806	1.50282	0.00524
33	bank	police	3.29	1.02105	0.43812	0.58293
34	bank	plumber	3.92	2.46380	2.12394	0.33986
35	bank	farmer	3.59	2.15061	1.82266	0.32795
36	bank	actor	3.72	2.17187	2.07121	0.10066

[결과 3.3]의 OBS는 디폴트로 제공되며 현재 9가지 직업을 각각 짝지었을 때 경우의 수 36가지를 나타내고 있고 이들의 짝을 _ROW_와 _COL_에서 볼 수 있다. 나머지 DATA, FITDATA, FITDIST, RESIDUAL을 [알고리즘 3]의 단계별로 비교하여 보면 DATA의 값은 [1단계]의 〈표 3.1〉의 비유사성 d_{ij}이고, FITDATA의 값은 [3단계]의 순위상 $\hat{d}_{ij}$이며, FITDIST의 값은 [2단계]에서 2-차원 형상좌표로부터 얻어지는 비유사성 δ^*_{ij}에 해당한다.

RESIDUAL의 값은 [4단계]의 스트레스 식 (3.3)의 분자에서 $-(\delta_{ij}^{*}-\hat{d}_{ij})$인 -잔차이다.

예를 들어 DATA의 첫 번째 행의 3.06은 〈표 3.1〉에서 두 개체 surgeon과 preach 사이의 비유사성을 나타낸다. FITDIST는 [결과 3.2]-①의 (b) 형상좌표로부터 얻어진다. 이를 [알고리즘 3]-[2단계]에서 형상좌표행렬을 $X^{*}=(x_{ij}^{*})$라 하고 이 행렬로부터 얻어지는 비유사성행렬을 $D^{*}=(\delta_{ij}^{*})$라고 하였다. 실제 FITDIS의 첫 번째 행인 $1.44773 \simeq 1.45$는 비유사성 $\delta_{surgeon,\,preach}^{*}$로 [결과 3.2]-①의 (b) 형상좌표로부터 다음과 같이 계산된다.

$$\begin{aligned}\delta_{surgeon,\,preach}^{*} &= \left[\sum_{k=1}^{2}(x_{surgeon,\,k}^{*}-\ x_{preach,\,k}^{*})^{2}\ \right]^{1/2}\\ &= [(-0.38-0.68)^{2}+\ (1.84-0.85)^{2}\ ]^{1/2}=\ 1.4504 \simeq 1.45.\end{aligned}$$

이번에는 FITDIST의 36번째 행의 $2.07121 \simeq 2.07$은 비유사성 $\delta_{bank,\,actor}^{*}$로

$$\begin{aligned}\delta_{bank,\,actor}^{*} &= \left[\sum_{k=1}^{2}(x_{bank,\,k}^{*}-\ x_{actor,\,k}^{*})^{2}\ \right]^{1/2}\\ &= [(-0.09-1.50)^{2}+\ (-1.07-0.25)^{2}\ ]^{1/2}=\ 2.0665 \simeq 2.07.\end{aligned}$$

이다.

이번에는 FITDATA와 FITDIST의 값은 순위상 $\hat{d}_{ij}$과 형상좌표의 비유사성 δ_{ij}^{*}로 나타내며 이들로부터 [결과 3.2]-③의 (a) 상그림을 얻을 수 있다. 상그림은 수직축의 순위상 $\hat{d}_{ij}$과 수평축의 비유사성 δ_{ij}^{*}에 대한 각 개체의 36가지 짝의 산점도이다. 이는 [프로그램 3.2]-1)과 2)에 의해서 자동으로 주어진다고 이미 언급한 바 있다. 이 그림에서 가장 이상적으로는 좌표점들이 원점을 지나는 직선상에 놓이면 비계량형 다차원척도법을 위한 순서척도모형이 적합함을 나타내게 된다. 직선상을 벗어나는 좌표점(이상치)들이 많을수록 식 (3.3)의 스트레스는 커지게 된다. 실제 이 스트레스는 FITDATA(순위상 $\hat{d}_{ij}$), FITDIST(비유사성 δ_{ij}^{*}), RESIDUAL(-잔차 $-(\delta_{ij}^{*}-\hat{d}_{ij})$)로부터 계산될 수 있다. 먼저 스트레스 식 (3.3)에서 분자는

$$\sum_{i<j}\sum(\delta_{ij}^{*}-\widehat{d_{ij}})^2 = \sum(\text{FITDIST}-\text{FITDATA})^2 \tag{3.4}$$
$$= \sum(-\text{RESIDUAL})^2$$
$$= (0.42668)^2+(0.28433)^2+\cdots+(-0.10066)^2$$
$$= 1.5372.$$

이고 분모는

$$\sum_{i<j}\sum(\delta_{ij}^{*})^2 = \sum(\text{FITDIST})^2 \tag{3.5}$$
$$= (1.44773)^2+(1.12308)^2+\cdots+(2.07121)^2$$
$$= 141.22984.$$

이다. 따라서 식 (3.4)와 식 (3.5)의 결과로부터 식 (3.3)의 스트레스 $ST1$은

$$ST1 = (1.5372/141.22984)^{1/2} = 0.1043283 \simeq 0.10$$

이다. 이는 [결과 3.2]-②의 〈Badness-of-Fit Criterion〉값과 일치하게 된다.

끝으로 [결과 3.2]-③의 (b) 세퍼드그림은 [프로그램 3.2]-3)의 결과로 수직축의 FITDATA (순위상 $\hat{d}_{ij}$)와 DATA(비유사성 d_{ij})에 의한 산점도이다. 이는 순서척도모형에서 가정한 $f(\cdot)$의 단조성을 평가하는데 실제로 본 사례에서 가정한 순서척도모형이 단조성을 보이고 있음을 알 수 있다.

특별히 Davison(1992, Chapter 5)은 비계량형 다차원척도법에 관한 이론 및 다양한 예제와 분석을 제공하고 있으니 관심 있는 독자는 참고 바란다.

3.3 계량형과 비계량형 다차원척도법의 비교

일반적으로 [알고리즘 2: 토거선 알고리즘]에 바탕을 두는 계량형 다차원척도법이 [알고리즘 3: 크루스칼-세퍼드 알고리즘]에 바탕을 두는 비계량형 다차원척도법보다는 수치적으로 계산이 더 간단하다. 그러나 전자가 거리함수의 단조 변환에 대하여 얼마나 로버스트한

것인지는 알려져 있지 않다. 다시 말해서 거리측정에 대한 주관적 판단을 사용하기보다는 비계량형 다차원척도법과 같이 비유사성의 순위순서를 유지하도록 함수 $f(\cdot)$를 객관적으로 세우는 것이다. 더군다나 두 방법은 잘 알려진 사례의 경우에는 유사한 결과를 제공한다. 이미 비계량형 다차원척도법이 적용된 1.1.2절 [사례 2] 우리나라 14개 경제관련기관의 경제 전망에 대하여 3.1.3절에서는 계량형 다차원척도법을 적용하고 있는데 그 결과는 대동소이함을 이미 살펴보았다. 이 절에서는 SAS/PROC MDS와 R-프로그램에서 제공하는 계량형과 비계량 다차원척도법의 프로그램과 결과를 비교하고자 한다.

3.3.1 SAS / MDS에서 비교

먼저 SAS/MDS에서 제공하는 두 방법의 비교를 위해 잘 알려진 사례를 하나 고려한다. Mardia 외 2인(1979, Chapter 14)에서는 점(. dot)과 선(- dash)의 짧은 신호로 구성된 모스부호(Morse code)에서 0, 1, 2, 3, 4, 5, 6, 7, 8, 9로 이루어진 10개의 아라비아 숫자에 대한 〈표 3.2〉와 같은 유사성자료를 제공하고 있다. 이 자료는 모스부호를 모르는 598명에게 청각적으로 기계에 의해서 두 수의 모스부호신호를 짝을 지어 두 번(1 다음 2, 2 다음 1과 같이)들려 주고 두 신호의 짝이 같다고 생각하는 사람들의 비율을 나타낸 비대칭 유사성행렬 자료 〈표 2.7〉의 하삼각행렬만 고려한 것이다. Seber(1984, Example 5.5)와 Izenman(2008, pp. 469-471)에서도 전체 모스부호자료에 대한 다차원척도그림을 제공하고 있으니 참고 바란다.

〈표 3.2〉 아라비아 숫자에 대한 모스부호의 유사성자료

1.----	84									
2..---	62	89								
3...--	16	59	86							
4....-	6	23	38	89						
5.....	12	8	27	56	90					
6-....	12	14	33	34	30	86				
7--...	20	25	17	24	18	65	85			
8---..	37	25	16	13	10	22	65	88		
9----.	57	28	9	7	5	8	31	58	91	
0-----	52	18	9	7	5	18	15	39	79	94

〈표 3.2〉의 아라비아 숫자에 대한 모스부호자료가 유사성행렬 $C=(c_{ij}),\ i,j=1,\ldots,10$를 이루고 이를 통해 비유사성행렬 $D=(d_{ij}),\ i,j=1,\ldots,10$를 얻기 위해 2.3절의 식 (2.9)의 표준변환

$$d_{ij}=\sqrt{c_{ii}-2c_{ij}+c_{jj}}$$

를 활용한다. 이것의 행렬표현은 다음과 같다.

$$D=\sqrt{diag(C)J-2C+Jdiag(C)}$$

여기서, $diag(C)$는 크기가 10×10 행렬 C의 대각원소로 이루어지는 대각행렬을 나타내며, J는 크기가 10×10이고 모든 원소가 1인 행렬이다.

[프로그램 3.3] 모스부호의 유사성자료의 계량형과 비계량형 다차원척도법을 위한 프로그램

```
data morsecode;
input code$  x1-x10;
cards;
 1.----      84   62   16    6    2   12   20   37   57   52
 2..---      62   89   59   23    8   14   25   25   28   18
 3...--      16   59   86   38   27   33   17   16    9    9
 4....-       6   23   38   89   56   34   24   13    7    7
 5.....      12    8   27   56   90   30   18   10    5    5
 6-....      12   14   33   34   30   86   65   22    8   18
 7--...      20   25   17   24   18   65   85   65   31   15
 8---..      37   25   16   13   10   22   65   88   58   39
 9----.      57   28    9    7    5    8   31   58   91   79
 0-----      52   18    9    7    5   18   15   39   79   94
;
proc iml;
   reset nolog;

  use morsecode;
  read  all var{code} into code;
  read all var{x1 x2 x3 x4 x5 x6 x7 x8 x9 x10} into C;
  n=nrow(C);
  J=J(n,n,1);
  /* Standard Transformation from C to D */
  D=sqrt(diag(C)*J -2*C+J*diag(C)); /* d_ij = sqrt[c_ii -2c_ij + c_jj] */
 print, "Dissimilarity matrix from Similarity matrix",
                D[rowname=code colname=code format=5.3];
```

```
  create dism from D[rowname=code];
  append from D[rowname=code];
  close dism;
  quit;

ods graphics on;
/* Metric MDS */
proc mds data=dism level=absolute pfinal;
      id code;

/* Non-metric MDS */
proc mds data=dism level=ordinal pfinal;
      id code;
run;
```

이 변환을 반영하여 [프로그램 3.3]은 /* Standard Transformation from C to D */에서 아라비아 숫자에 대한 모스부호의 유사성자료 〈표 3.2〉로부터 비유사성행렬을 계산하고 이에 대하여 계량형 다차원척도법과 비계량형 다차원척도법을 하기 위한 대한 PROC MDS 프로그램이다. 계량형 다차원척도법은 /* Metric MDS */에서 옵션 〈level=absolute〉

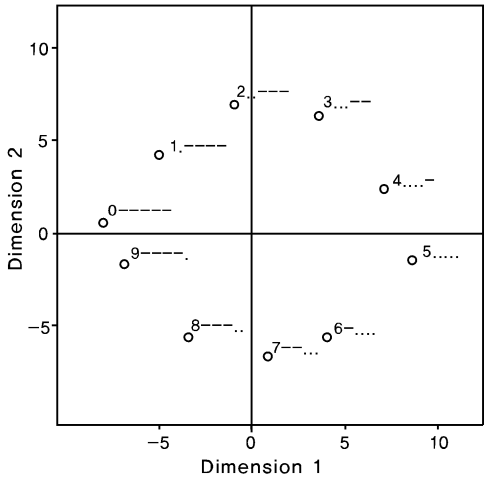
로, 비계량형 다차원척도법은 /* Non-metric MDS */ 옵션 〈level=ordinal〉로 이루어진다.

이 [프로그램 3.3]의 수행 결과들 중 비유사성행렬과 다차원척도그림만 [결과 3.4]에서 제공하고 있다. 결과에는 요약되지 않았지만 적합도는 (a) 계량형 다차원척도그림이 0.21이고 (b) 비계량형 다차원척도그림이 0.05으로 〈표 1.9〉의 크루스칼의 판별기준에 따르면 각각 나쁨(poor)와 좋음(good)에 해당하여 상대적으로 계량형 다차원척도그림보다는 비계량형 다차원척도그림이 적합도에서는 좋다고 여겨진다. 그러나 두 그림의 경향은 대동소이하다. 전체적으로 1축의 왼편에서부터 위쪽으로 0(-----), 1(.----), 2(..---), 3(...--), 4(....-), 5(.....), 6(-....), 7(--...), 8(---..), 9(----.) 순서로 다소 거칠지만 원을 형성하고 있다. 수평축(1차원축)은 왼쪽에서 오른쪽으로 갈수록 모스부호인 점(.)의 증가되는 추세를 측정하며, 그 반면에 수직축(2차원축)은 위쪽은 아라비아 숫자 1(.----), 2(..---), 3(...--), 4(....-)로 점(.)이 먼저 시작되며 아래쪽은 9(----.), 8(---..), 7(--...), 6(-....)으로 선(-)이 먼저 시작되어 신호의 동질성을 측정한다. 0과 5는 2차원축의 이런 경향에서 벗어나 있다. 이를 제대로 분류하게 위해서는 좌표축을 시계방향으로 (a)에서는 30도 가량, (b)에서는 20도 가량 회전할 필요가 있다. 회전한 후에 특히, (a) 계량형 다차원척도그림에서는 좌표점을 나타내는 숫자를 원점에서 시작되는 벡터로 보면 0과 9, 그리고 4와 5는 각각 그들이 서로 이루는 각이 매우 작아 매우 연관성이 높게 여겨진다. 그 이유로는 마지막 한 개의 신호에서 차이를 보여 실험에 참가한 모스부호에 익숙하지 않은 사람들이 그 차이를 인지하지 못하기 때문이라 여겨진다. 그리고 이러한 기하적인 해석은 (a) 계량형 다차원척도그림이 주성분분석의 대수적인 면과 기하적인 면을 따르는 [알고리즘 2: 토거선 알고리즘]에 바탕을 두고 있고 주성분행렬도(principal component biplot)의 관점에서 이런 기하적인 해석이 유용하게 사용되었다. 행렬도에 대해서는 최용석과 신상민(2013, 1장)을 참고 바란다.

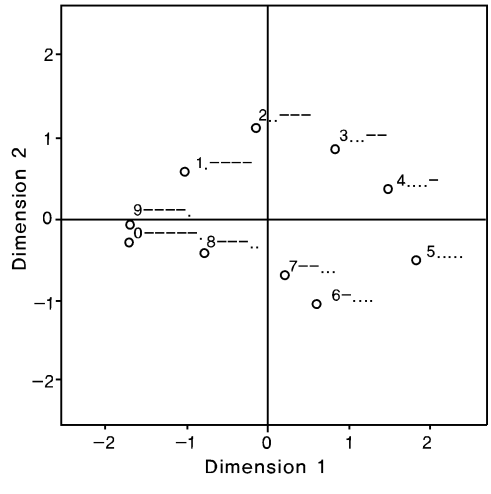
[결과 3.4] [프로그램 3.3]의 실행 결과 중 비유사성행렬과 다차원척도그림

Dissimilarity matrix from Similarity matrix

	CM									
	1.----	**2..---**	**3...--**	**4....-**	**5.....**	**6-....**	**7--...**	**8---..**	**9----.**	**0-----**
1.----	0.000	7.000	11.75	12.69	13.04	12.08	11.36	9.899	7.810	8.602
2..---	7.000	0.000	7.550	11.49	12.77	12.12	11.14	11.27	11.14	12.12
3...--	11.75	7.550	0.000	9.950	11.05	10.30	11.70	11.92	12.61	12.73
4....-	12.69	11.49	9.950	0.000	8.185	10.34	11.22	12.29	12.88	13.00
5.....	13.04	12.77	11.05	8.185	0.000	10.77	11.79	12.57	13.08	13.19
6-....	12.08	12.12	10.30	10.34	10.77	0.000	6.403	11.40	12.69	12.00
7--...	11.36	11.14	11.70	11.22	11.79	6.403	0.000	6.557	10.68	12.21
8---..	9.899	11.27	11.92	12.29	12.57	11.40	6.557	0.000	7.937	10.20
9----.	7.810	11.14	12.61	12.88	13.08	12.69	10.68	7.937	0.000	5.196
0-----	8.602	12.12	12.73	13.00	13.19	12.00	12.21	10.20	5.196	0.000



(a) 계량형 다차원척도그림



(b) 비계량형 다차원척도그림

3.3.2 R-프로그램에서 비교

R-프로그램에서는 계량형과 비계량형 다차원척도법을 위한 프로그램은 cmdscale 함수와 isoMDS 함수가 각각 제공된다. 특히, isoMDS 함수를 이용하기 위해선 R-프로그램에서 라이브러리 MASS를 지정해야 한다.

[프로그램 3.4] 모스부호의 비유사성자료의 계량형과 비계량형 다차원척도법을 위한 R-프로그램

```
# Metric MDS
d<- as.matrix(read.table("c://R-MDS//morse.txt", header=T, check.names=F))
con<- cmdscale(d, k=2, eig=T)
con
x<-con$points[,1]
y<-con$points[,2]
lim<-c(-max(abs(con$points)), max(abs(con$points)))
plot(x,y, xlab="Dimension 1", ylab="Dimension 2", xlim=lim, ylim=lim)
text(x,y+0.6, colnames(d), cex=0.8)
abline(v=0, h=0)

# Nonmetric MDS
d<- as.matrix(read.table("c://R-MDS//morse.txt", header=T, check.names=F))
library(MASS)
con<- isoMDS(d, k=2)
con
x<-con$points[,1]
y<-con$points[,2]
lim<-c(-max(abs(con$points)), max(abs(con$points)))
plot(x,y, xlab="Dimension 1", ylab="Dimension 2", xlim=lim, ylim=lim)
text(x,y+0.6, colnames(d), cex=0.8)
abline(v=0, h=0)
```

3.3.1절에서 살펴본 〈표 3.2〉 아라비아 숫자에 대한 모스부호의 유사성자료로부터 표준변환 $d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}}$ 에 의한 비유사성행렬자료에 대하여 계량형과 비계량형 다차원척도법에 관한 R-프로그램이 [프로그램 3.4]이다. R-프로그램에서 다차원척도법을 위한 함수에 사용되는 입력자료는 반드시 비유사성행렬자료이어야 한다.

[프로그램 3.4]의 # Metric MDS 이하는 계량형 다차원척도법에 관한 cmdscale 함수를 이용하였고, # Nonmetric MDS 이하에는 비계량형 다차원척도법을 위한 isoMDS 함수를 사용하고 library(MASS) 문에 의해서 라이브러리 MASS를 지정하였다. 공통적으로 먼저 다음의 명령문

```
d<-as.matrix(read.table("c://R-MDS//morse.txt",
                                    header=T, check.names=F))
```

에서 read.table 함수의 〈c://R-MDS//morse.txt〉는 드라이브 c:에 위치한 폴더 〈R-MDS〉의 자료 〈morse.txt〉를 불러들여 행렬 〈d〉로 저장하라는 명령문이다. 〈header=T, check.names=F〉는 첫 행에 개체인 10개의 아라비아 숫자와 모스부호를 나타내는 1(.----), 2(..---), 3(...--), 4(....-), 5(.....), 6(-....), 7(--...), 8(---..), 9(----.), 0(-----.)의 문자자료가 있음을 알리는 인수이다.

cmdscale과 isoMDS 함수에서 $k=2$는 2차원의 형상공간을 적합하라는 인수(argument)이다. 특히, isoMDS 함수에는 알고리즘에서 초깃값으로 계량형 다차원척도법의 형상좌표를 이용하기 위하여 y=cmdscale(d,k)에 의한 cmdscale 함수를 이용하는 것이 디폴트 인수로 주어진다. 더불어 수치적 반복을 위해 niter=100과 수렴의 조건 tol=1e-4 역시 디폴트 인수로 주어진다.

[프로그램 3.4]에서 실제 두 방법의 비교를 위해 3.3.1절에서 활용한 자료를 고려한다. 즉, 점(. dot)과 선(- dash)의 짧은 신호로 구성된 모스부호(Morse code)에서 0, 1, 2, 3, 4, 5, 6, 7, 8, 9로 이루어진 10개의 아라비아 숫자에 대한 〈표 3.2〉의 유사성자료이다. 이는 3.3.1절의 [프로그램 3.3]의 수행결과인 비유사성행렬자료를 [프로그램 3.4]에서 입력자료로 활용되었다. [부록 2: 자료 10(파일이름: morse.txt)]를 참고 바란다.

[프로그램 3.4]의 수행 결과들인 (a) 계량형 다차원척도법과 (b) 비계량형 다차원척도법의 ① 형상좌표 ② 최종적합도 ③ 다차원척도그림이 [결과 3.5]에 요약 정리되어 있다. R-프로그램에서 다차원척도법의 최종 적합도는 계량형의 경우 식 (3.1)의 적합도 $\alpha_{1,k}$를 사용하되 퍼센트(%)가 아닌 형태로 실제 결과값에 100을 곱해야 한다. 다음으로 비계량형인 경우 퍼센트(%) 형태로 제공되므로 그 차이를 잘 인식하기를 바란다. (a) 계량형 다차원척도그림이 0.56(56%) (b) 비계량형 다차원척도그림이 4.83%(0.05)이다. 물론 (a) 계량형 다차원척도그림의 적합도 0.56은 [결과 3.4]의 적합도에서 0.21에 비해 다소 높아졌지만 둘 다 적합도에서는 좋지는 않다. 전반적으로 다차원척도그림에 대한 해석은 3.3.1절의 SAS/MDS의 [결과 3.4]와 거의 동일하나 약간의 차이점이 있다. 가장 두드러진 차이점은 2차원축(Dimension 2)에 대하여 위쪽과 아래쪽 좌표점의 위치가 기하적으로 바뀌어 있지만 해석상 전혀 문제가 되지 않으며 0(-----)이 9(----.)쪽으로 이동해 있다는 것이다. 전반적으로 1차원축의 왼편으로부터 아래쪽으로 0(-----), 1(.----), 2(..---), 3(...--), 4(....-), 5(.....), 6(-....), 7(--...), 8(---..), 9(----.) 순서로 다소 거칠지만 원을 형성하고 있다. 수평축(1차원축)은 왼쪽에서 오른쪽으로 갈수록 모스부호인 점(.)

의 증가되는 추세를 측정하며 그 반면에 수직축(2차원축)의 아래쪽은 1(.----), 2(..---), 3(...--), 4(....-)로 점(.)이 먼저 시작되며 위쪽은 0(-----), 6(-....), 7(--...), 8(---..), 9(----.)로 선(-)이 먼저 시작되어 신호의 동질성을 측정한다. 다만 5(.....)는 이 경향에서 벗어나 있다. 반면에 이런 경향은 SAS/MDS의 경우에는 0(-----)도 해당되었던 것이 다르다. R-프로그램의 [결과 3.5]에서는 두 축을 시계 반대 방향으로 10도 정도 회전하면 2차원축의 해석이 명확해진다.

[결과 3.5] [프로그램 3.4]의 실행 결과

(a) 계량형 다차원척도그림	(b) 비계량형 다차원척도그림

①

```
$points
           [,1]       [,2]
 [1,] -4.5711897 -3.1336272
 [2,] -0.4107332 -5.9656034
 [3,]  3.6367506 -4.5046057
 [4,]  5.6351188 -0.6246504
 [5,]  5.6460359  0.4240691
 [6,]  3.7437600  3.7077844
 [7,]  0.9471871  5.2112558
 [8,] -2.9273268  3.6211882
 [9,] -6.0972424  0.8105425
[10,] -5.6023603  0.4536467
```

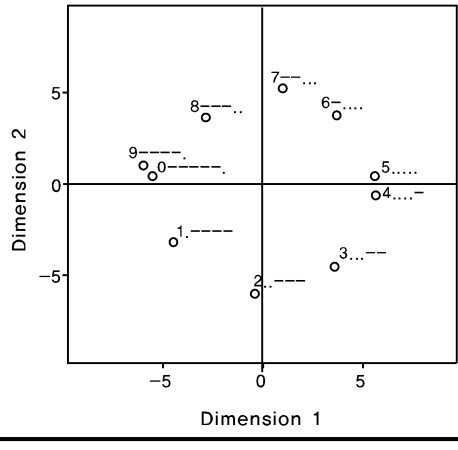
```
$points
           [,1]      [,2]
 [1,] -4.4095078 -2.123888
 [2,] -0.7155983 -4.529155
 [3,]  3.1874478 -3.727998
 [4,]  6.6531606 -2.029248
 [5,]  9.0031730  1.179337
 [6,]  2.7541260  4.226263
 [7,]  0.9314980  3.025961
 [8,] -3.1950315  2.018792
 [9,] -7.0021477  0.501086
[10,] -7.2071202  1.458849
```

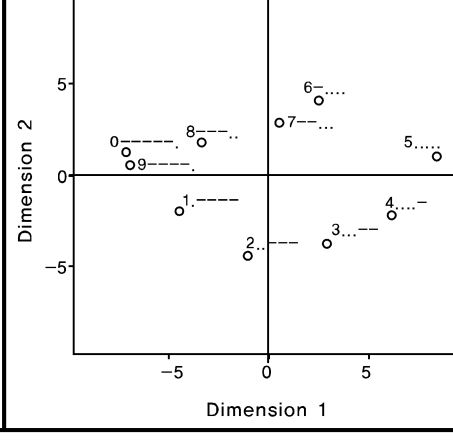
②

```
$GOF
[1] 0.5583503 0.5622701
```

```
$stress
[1] 4.83407
```

③





연습문제

3.1 전국 17개 주요 은행별 경영평가를 8가지 항목에 대해 정리한 자료이다.

		경영평가 항목							
구분	은행	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
인수되는 은행	동화	0	0	0	0	0	1	1	0
	동남	0	0	0	0	1	1	0	0
	대동	0	0	0	0	0	0	1	0
	충청	0	0	0	0	0	1	0	1
	경기	0	0	0	0	0	1	1	1
인수하는 은행	신한	1	1	1	1	1	1	1	0
	주택	1	1	1	1	1	1	1	0
	국민	1	1	1	1	1	1	1	1
	하나	1	1	0	1	1	1	1	1
	한미	1	1	0	1	1	1	1	1
경영평가 대상은행	조흥	0	1	1	0	0	1	1	1
	상업	0	1	1	0	0	1	1	1
	한일	0	1	1	1	0	1	1	1
	외환	0	1	1	0	0	1	1	1
	평화	0	0	0	0	1	0	1	1
	강원	0	0	0	0	0	0	0	1
	충북	0	0	0	0	1	0	0	0

이 자료는 중앙일보(1998.6.29)에서 퇴출된 은행 및 시중 17개 은행에 대한 경영평가를 조사한 계량치 자료를 다음의 8가지 항목을 기준으로 이진수자료로 가공한 것이다.

(1) 국제결제은행(BIS)가 자기자본이 8% 이상인가?
(2) 총자산이 130,000억 원 이상인가?
(3) 무수익 여신이 8,000억 원 이상인가?
(4) 총여신대비 무수익 여신비율이 6% 이하인가?
(5) 부실여신은 3,300억 원 이하인가?
(6) 점포수는 110개 이상인가?
(7) 인원수는 1700명 이상인가?
(8) 중소기업 지원 성적이 B 이상인가?

1) 단순매칭계수와 러셀-라오계수를 계산하여 보라.
2) 비유사성행렬자료를 얻어 보라.
3) 토거선 알고리즘의 단계별로 계산 결과를 제시하라.
4) 계량형과 비계량형 다차원척도그림을 얻고 해석하고 비교하라.
5) 적합도에 대하여 논의하라.
6) R-프로그램을 활용하여 계량형과 비계량형 다차원척도법을 수행하라.
그리고 SAS/MDS 결과와 비교하라.

3.2 다음은 우리나라 10대 도시별 지도상의 거리(cm)를 나타낸 자료이다(문교부, 1972, 대한민국전도(척도 1:1,000,000)). 계량형과 비계량형 다차원척도법을 모두 실시하여 서로 비교하여라.

서울	.									
부산	32.5	.								
광주	27	19.5	.							
대구	24.2	9.5	17	.						
인천	4	34.4	16.5	25	.					
강릉	17	29.6	35.5	21	20.7	.				
청주	11.7	22	17.5	13.5	11.9	18	.			
평양	20	52	44	43	19.4	30.7	31	.		
신의주	30.6	68.5	60	58.9	30.4	47.2	46	17	.	
전주	19.2	18	7.6	12.8	19.2	26.5	9.8	37.8	52	.

3.3 다음은 6가지 종목의 운동경기에 대한 비유사성을 나타낸 자료이다(Davison, 1992, Table 4.1). 이는 경기의 속도감과 선수들 간의 신체적 접촉의 정도에 따라 얼마나 다른지를 보여 주고 있다. 예를 들면 하키와 축구는 매우 빠르면서 신체적 접촉이 있는 운동이고 골프와 크로케는 느리면서 신체적 접촉이 없는 경기이다.

하키	0.00					
축구	0.71	0.00				
농구	1.41	1.41	0.00			
테니스	1.73	1.73	1.00	0.00		
골프	2.00	2.00	1.41	1.00	0.00	
크로케	2.00	2.00	1.41	1.00	0.71	0.00

1) 계량형과 비계량형 다차원척도그림을 얻고 해석하라.
2) 운동경기의 유형을 다차원척도그림에서 찾아라.
3) R-프로그램을 활용하여 계량형과 비계량형 다차원척도법을 수행하라.

3.4 다음은 미국의 10대 도시 간의 비행노선의 거리(마일)를 제일 짧은 노선을 순위 1(New York과 Washington D.C.)로 하고 제일 긴 노선을 순위 45(Miami와 Seattle)로 가공한 순위자료이다(SPSS Inc., 1992, p. 173).

Atlanta	.									
Chicago	4	.								
Denver	22	13	.							
Houston	8	15	12	.						
Losangel	34	31	11	24	.					
Miami	6	21	29	18	39	.				
Newyork	10	9	27	25	42	20	.			
Sanfran	35	32	16	28	2	44	43	.		
Seattle	36	30	19	33	17	45	40	7	.	
Washdc	3	5	26	23	37	14	1	41	38	.

1) 비계량형 다차원척도법을 실시하라.
2) [문제 3.1]에서 구한 비유사성행렬을 위와 같이 순위자료로 만들고 비계량형 다차원척도법을 실시하라. [문제 3.1]의 비계량형 다차원척도법의 결과와 비교할 때 어떤 차이가 있는지 설명하라.

3.5 미국 내 22개 공익회사에 대한 평가를 위해 8개 변수로 측정한 자료이다(최용석, 정광모, 2003, 5장). 여기서 8개 변수는 X_1: 고정요금 보상율, X_2: 자금 수익률, X_3: KW당 비용, X_4: 연부하율, X_5: 1974-1975년 사이 절정의 kWh 수요성장, X_6: 판

매량, X_7: 핵비율, X_8: 총연료비용이다. [부록 2: 자료 7(파일이름: utility.txt)]을 참고하라.

공익회사	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1. 애리조나	1.06	9.2	151	54.4	1.6	9077	0	0.628
2. 보수톤에디슨	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
3. 루이지내나	1.43	15.4	113	53.0	3.4	9212	0	1.058
4. 코몬웰스에디슨	1.02	11.2	168	56.0	0.3	6423	34.3	0.700
5. 콘솔에디슨	1.49	8.8	192	51.2	1.0	3300	15.6	2.044
6. 플로리다파우워	1.32	13.5	111	60.0	−2.2	11127	22.5	1.241
7. 하와이안일렉트릭	1.22	12.2	175	67.6	2.2	7642	0	1.652
8. 아이다호 파워	1.10	9.2	245	57.0	3.3	13082	0	0.309
9. 켄터키유틸	1.34	13.0	168	60.4	7.2	8406	0	0.862
10. 매디슨가스	1.12	12.4	197	53.0	2.7	6455	39.2	0.623
11. 네바다파워	0.75	7.5	173	51.5	6.5	17441	0	0.768
12. 뉴잉글랜드파워	1.13	10.9	178	62.0	3.7	6154	0	1.897
13. 노던스테이트파워	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
14. 오클라호마가스	1.09	12.0	96	49.8	1.4	9673	0	0.588
15. 퍼시픽가스	0.96	7.6	164	62.2	−0.1	6468	0.9	1.400
16. 푸젯싸운드파워	1.16	9.9	252	56.0	9.2	15991	0	0.620
17. 샌디에고가스	0.76	6.4	136	61.9	9.0	5714	8.3	1.920
18. 서던캄퍼니	1.05	12.6	150	56.7	2.7	10140	0	1.108
19. 텍사스유틸	1.16	11.7	104	54.0	−2.1	13507	0	0.636
20. 위스콘신일렉트릭	1.20	11.8	148	59.9	3.5	7287	41.1	0.702
21. 유나이트일루미네이팅	1.04	8.6	204	61.0	3.5	6650	0	2.116
22. 버지니아일렉트릭	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

1) 비유사성행렬로 유클리드거리와 마할라노비스거리를 구하라.
2) 두 거리의 다차원척도법을 위한 적절한 차원수와 적합도를 구하라.
3) 1)에서 구한 각각의 거리에 대한 다차원척도그림을 통해 적절한 군집과 특성을 해석하고 비교하라.
4) R-프로그램을 활용하여 계량형과 비계량형 다차원척도법을 실시하라.

3.6 Mardia 외 2인(1979, pp. 412-413)은 다음과 같이 정의되는 유사성행렬 $C=(c_{ij})$를 소개하고 있다. 이를 켄달(Kendall)의 행렬이라 부르며 편자효과(horseshoe effect)를 보여 주기 위한 가공자료이다.

$$c_{ii} = 9,\ c_{ij} = \begin{cases} 8,\ 1 \le |i-j| \le 3, \\ \quad \cdots, \\ 1,\ 22 \le |i-j| \le 24, \\ 0,\ |i-j| \ge 25. \end{cases}$$

1) 편자효과에 대하여 기술하고 이 유사성 행렬의 크기를 말하라.
2) 표준변환을 통해 비유사성행렬 $D = (d_{ij})$를 구하라.
3) R-프로그램을 활용하여 계량형과 비계량형 다차원척도법을 실시하라.
4) 3)에서 구한 다차원척도그림을 통해 편자효과가 있음을 설명하라.

3.7 22명의 남자(M)와 18명의 여자(F)의 두개골에 대한 12개 측정변수에 의해 측정된 자료가 다음과 같이 주어져 있다(Cox와 Cox, 1994, pp. 30-34). 이 자료는 고대 이집트 Naqada 종족의 두개골 자료에 대한 칼피어슨(Karl Pearson)의 연구에 의해서 주어졌고 [부록 2: 자료 11(파일이름: skull.txt)]에 제공되어 있다. 12개 측정변수는 다음과 같다.

1) Greatest length(L), 2) Breadth(B), 3) Height(H), 4) Auricular height(OH), 5) Circumference above the superciliary ridger(U), 6) Sagittal circumference(S), 7) Cross-circumference(Q), 8) Upper face height(FH), 9) Nasal breadth(NB), 10) Nasal height(NH), 11) Cephalic index(B/L), 12) Ratio of height to length(H/L).

	L	B	H	OH	U	S	Q	FH	NB	NH	B/L	H/L
5F	−2.2700	−0.7810	−1.7400	−1.3300	−1.8900	−0.9020	−1.4800	−1.9000	−1.4100	−1.0600	1.4120	0.4960
7M	0.0364	0.3764	−0.9660	−0.6850	0.3166	−0.6520	−0.9220	0.1855	−0.4530	0.4895	0.2161	−1.1200
10F	−0.7070	1.5570	−0.3270	0.1476	−0.4180	−0.4010	1.1240	0.5110	0.6635	−0.8000	1.7180	0.3807
13M	0.9280	1.4390	−0.6750	−0.0960	1.2610	0.1838	1.5570	1.3570	2.5770	0.7475	0.2161	−1.6600
26M	1.6710	1.6760	1.1630	1.6210	2.1710	1.5200	2.0530	0.4459	0.9187	0.2315	−0.2130	−0.4650
32M	−0.5580	0.8488	−1.9300	−1.0700	−0.2780	−1.4000	−0.3640	−0.8560	−0.2930	1.2630	1.0740	−1.5800
43F	−1.4500	0.2583	−1.5500	−1.0100	−1.0100	−1.1500	−0.4880	−0.5960	−1.0900	−0.0260	1.4420	−0.1960
45F	−0.4090	−0.4500	−1.4500	−2.6300	−1.0500	−1.5700	−1.1100	−0.5960	−0.7720	−2.0900	0.0015	−1.2000
46F	−1.3800	1.0850	−0.6750	−0.1730	−0.2080	−0.4010	0.8756	−1.7700	−0.4530	−0.2840	1.9940	0.7268
52M	−0.5580	−0.0960	−0.5790	−0.1730	−0.3130	−0.4010	−0.1160	0.4459	−0.1340	0.7475	0.3694	−0.0810
58F	0.2593	−1.0400	0.1955	0.1476	0.3516	0.1838	−0.2400	1.2270	1.1420	0.7475	−0.9790	−0.0810
59F	−0.1120	−2.0600	−0.7720	0.7241	−0.7330	−0.0670	−0.6120	−1.5100	−1.4100	0.2315	−1.4100	−0.7340
63F	0.5565	0.7307	0.4858	1.2360	0.8414	1.6040	1.2470	−1.2500	−0.7720	0.2315	0.0322	−0.0810
64F	−1.7500	−1.0400	−0.5790	−1.8400	−1.7500	−2.5300	−2.3500	−1.2500	−0.4530	−1.3200	0.5839	1.2650
66M	−0.4090	1.2030	−0.9660	0.0836	−0.0680	−1.4900	0.8756	1.2270	0.6635	2.2950	1.1970	−0.6580
70F	−1.8700	−0.9230	−0.3850	−0.4930	−1.9600	−1.1500	−1.1100	−1.1200	−2.3700	−0.0260	0.9211	1.6110
83F	1.0770	1.5570	−0.3850	0.2117	1.1910	1.1030	0.3797	0.4459	1.3810	−0.8000	0.1854	−1.5000
85M	0.8537	0.8488	0.1955	−0.5570	0.9464	0.3091	−0.3640	−1.6400	−0.7720	−0.8000	−0.1210	−0.6580
86M	−0.1860	0.0221	0.9696	0.3398	−0.0330	0.3091	0.1317	0.7062	−0.2930	−0.8000	0.1548	1.2650
93bM	−1.1500	0.6126	−1.3500	−1.0700	−0.5580	−1.2400	0.2557	−0.4650	0.6635	−1.0600	1.4420	−0.3110
96M	1.0020	1.3210	1.7440	1.7490	1.2610	1.7710	2.1150	−0.5960	−0.4530	0.7475	0.0935	0.8037
97F	−1.6000	1.0850	0.0019	0.1476	−1.3600	−0.7350	−0.1160	0.5761	0.6635	0.8764	2.2090	1.7650
99M	0.9280	0.3764	0.9696	1.3650	0.6665	0.6015	0.7516	−0.2050	−1.2500	−1.0600	−0.5200	0.0731
102M	0.1850	0.1402	0.1955	−0.8130	−0.4880	−0.1500	−0.4880	0.4459	1.1420	−1.3200	−0.0600	−0.0040
112M	0.4822	−2.2200	0.8342	−0.0450	−0.4880	−0.2340	−0.8600	0.6411	1.3020	−0.2840	−1.9600	0.3807
120M	2.1910	0.4945	0.9696	0.2117	2.4510	0.8521	0.5036	1.7480	1.3020	2.2950	−1.3800	−1.1600
121F	0.0364	−0.4500	−1.1600	−1.3300	−0.4180	−0.9020	−0.8600	−0.3350	−0.4530	−0.2840	−0.3660	−1.3100
125M	−0.2610	−0.6860	1.6470	1.2360	−0.3130	0.1838	0.9995	−0.8560	−0.6120	−1.4500	−0.3050	2.0730
136M	1.2250	−1.0400	−0.3850	0.7241	0.8414	1.1030	−0.1160	1.0970	0.5838	−0.0260	−1.7200	−1.6600
137F	0.1850	−0.9230	−0.3850	−0.5570	−0.2080	0.3091	−1.2300	1.4870	0.1053	0.2315	−0.8260	−0.6190
138M	0.3336	−0.3320	−0.1920	0.2117	0.0367	0.1838	−0.1160	−0.8560	−0.7720	0.7475	−0.5200	−0.5810
139M	0.0364	−0.9230	−0.3850	0.0836	−0.6980	−0.0670	−0.3640	−0.3350	0.5040	1.0050	−0.7030	−0.4650
140F	0.7794	0.0221	0.9696	1.7490	0.7714	0.9356	1.6190	0.7062	0.8230	−1.8300	−0.6420	0.2269
143M	0.1850	0.0221	0.7761	−0.4290	0.1067	0.1002	−0.1160	0.3157	−0.2140	−0.5420	−0.1520	0.6499
144F	1.0770	−0.5210	0.9696	1.4930	0.9114	1.1440	0.2557	0.8364	−0.7720	1.0050	−1.2200	−0.0810
145F	0.1850	−0.0960	1.7440	0.7241	0.1417	1.8550	0.5036	−0.0750	0.5040	0.6443	−0.2440	1.6880
146F	0.0364	−0.5090	0.3890	−1.3300	−0.9080	0.1002	−1.4800	0.7062	−0.7720	−0.1550	−0.4280	0.3807
148M	0.1850	−1.7500	0.2922	0.3398	0.0717	−0.1500	−0.4880	−0.8560	−0.7720	−0.0260	−1.4100	0.1115
151M	−1.0000	−0.8050	0.3890	0.0836	−0.6280	−0.1500	−0.8600	−0.0750	0.0255	0.2315	0.2468	1.4960
152M	1.2250	0.9669	1.9370	0.9803	1.4360	1.3950	0.9995	2.0080	1.7800	1.2630	−0.3360	0.8037

1) 개체들 간의 적절한 거리를 구하고 그 타당성을 설명하라.

2) 계량형 다차원척도법을 실시하고 남녀 군집에 차이를 설명하라.

3) 주성분분석의 주성분점수를 이용하여 1)의 결과와 비교하라.

4) 남녀 따로 분리된 자료에 대한 각각의 계량형 다차원척도법을 실시하라.

5) 남자 두개골에 대한 다차원척도그림에서 나머지 개체들과 다른 개체를 찾아보고 그

특징을 설명하라.

6) 여자 두개골에 대한 다차원척도그림에서 나머지 개체들과 다른 개체를 찾아보고 그 특징을 설명하라.

3.8 40명으로부터 24대 자동차(모델) 선호도 평가를 8가지 관점(X_1: 경제성, X_2: 서비스, X_3: 비감가상각, X_4: 가격(매우 저렴한 차는 1점), X_5: 디자인, X_6: 스포티카 X_7: 안전성, X_8: 다루기 쉬움)에서 7점 척도(1(매우 좋다) ~ 6(매우 나쁘다))에 의한 설문조사로부터 얻어진 평균점수가 [부록 2: 자료 13(파일이름: cartype.txt)]에 제공되어 있다(H rdle와 Simir, 2007).

1) 비유사성행렬로 유클리드거리와 마할라노비스거리를 구하라.
2) 두 거리의 계량형과 비계량 다차원척도법을 각각 실시하라.
3) 다차원척도그림을 통하여 자동차의 특성에 따른 적절한 군집이 얻어지는를 살펴보라.

4장
삼원 다차원척도법

3장의 이원 다차원척도법은 n개 개체들의 비유사성 d_{ij}, $i,j=1,\ldots,n$에 의한 이원(two-way)행렬 형태인 한 개의 비유사성행렬 D에 적용된 다차원척도법을 뜻한다. 더 나아가 이번에는 m명의 주체(subjects) 각각에 대해서 고려한 두 개 이상의 비유사성행렬 $D^{(s)}$, $s=1,\ldots,m$에 적용되는 삼원 다차원척도법을 살펴보기로 하자. 삼원(three-way)이란 이원행렬 형태의 비유사성행렬을 주체에 대하여 각각 고려한 것을 말한다. 이를 위한 INDSCAL(INdividual Differences SCALing) 모형의 알고리즘을 소개하고 기하적 관점에서 이해하려고 한다. 또한 실제 삼원 다차원척도법을 적용하고 이해하기 위해 사례를 통하여 프로그램과 수행 결과를 제시하려 한다.

4.1 삼원 다차원척도법을 위한 모형

삼원 다차원척도법을 위한 모형과 관련된 알고리즘으로 Tucker와 Messick(1963), Horan(1969) 그리고 Carroll과 Chang(1970)의 것이 있다. 가장 대표적인 방법이 Carroll과 Chang이 제안한 INDSCAL이다. 그들은 3.1.1의 [알고리즘 2: 토거선 알고리즘]에서 각 주체의 가중치를 고려한 모형을 제안하였고 이를 종종 INDSCAL 모형이라고 불렀다. Davison(1992, pp. 121-123)은 가중치를 차원가중치, 중요(importance)가중치 혹은 중요점(salience)가중치라 한다. 그러나 엄격히 말하자면 INDSCAL 모형은 삼원 다차원척도법을 제공하는 컴퓨터 프로그램을 일컫는다(Schiffman 외 2인, 1981, p. 348; Davison, 1992, p. 121). 특히 삼원 다차원척도법을 위한 INDSCAL 모형의 내용에 대해서는 Davison(1992, Chapter 6), Young(1987, Chapter 6) 그리고 Schiffman 외 2인(1981, pp. 371-389)을 참고할 수 있다.

4.1.1 INDSCAL 모형의 소개

이 절에서는 Schiffman 외 2인(1981, p. 372)이 정리 요약한 INDSCAL 모형을 소개하려고 한다. 이미 식 (1.2)에서 정의한 m명의 주체에 대해서 고려한 삼원행렬 형태인 비유사성행렬

$$D^{(s)} = (d_{ij}^{(s)}), \quad i,j = 1,\ldots,n;\ s = 1,\ldots,m$$

을 고려하자. 여기서 $d_{ij}^{(s)}$는 s번째 주체에 대해 i번째와 j번째 개체 사이의 가중유클리드 거리에 의한 비유사성을 나타낸다.

삼원 다차원척도법에서 궁극적으로 구하고자 하는 $k\,(\le p)$ 차원 형상공간에서 크기가 $n \times k$인 형상좌표를 $X_{(k)}$ 라 하자. s번째 주체에 대한 r번째 차원에 대한 가중치를 w_{sr}라고 하고, i번째와 j번째 개체 간의 비유사성을

$$\delta_{ij}^{(s)} = \left[\sum_{r=1}^{k} w_{sr} (x_{ir} - x_{jr})^2 \right]^{1/2} \tag{4.1}$$

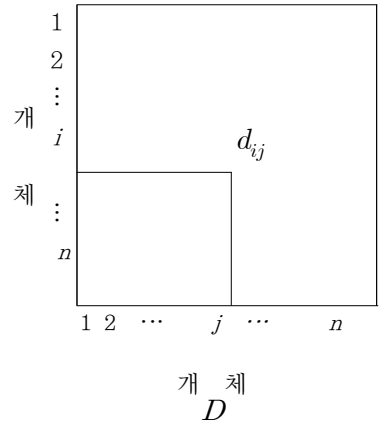
라고 하면 삼원 다차원척도법을 위한 모형은

$$d_{ij}^{(s)} \;=\; f_s(\delta_{ij}^{(s)}) + \epsilon_{ij} \tag{4.2}$$

를 만족한다. 여기서 $f_s(\cdot)$는 일반적으로 계량형인 경우 선형함수이고 비계량형인 경우는 단조함수이며 ϵ_{ij}는 왜곡오차이다.

입력: $n \times n$ 비유사성행렬 $D = (d_{ij})$.

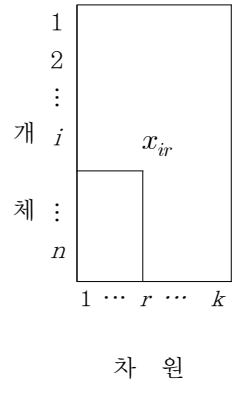
$d_{ij} \equiv$ 개체 i와 j의 비유사성.



결과: k차원의 $n \times k$ 형상좌표행렬 $X_{(k)}$.

$x_{ir} \equiv$ 차원 r에서 개체 i의 좌표.

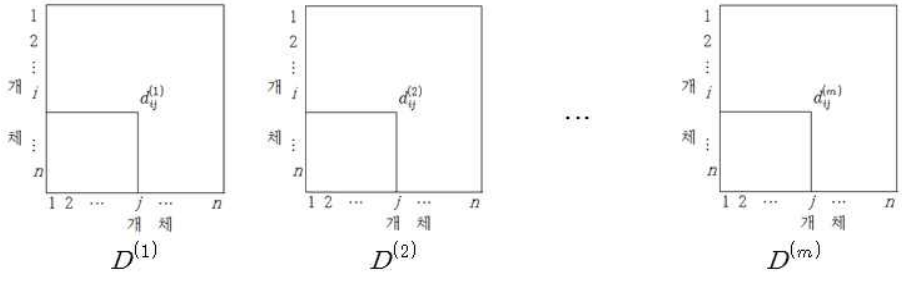
$\delta_{ij} \equiv$ 형상공간에서 개체 i와 j의 비유사성.



〈그림 4.1〉 이원 다차원척도법을 위한 입력자료와 결과의 도식도

입력: $m(\geq 2)$개의 $n \times n$ 비유사성행렬 $D^{(s)}$, $s=1,\ldots,m$.

$d_{ij}^{(s)} \equiv$ 주체 s에 대해 개체 i와 j의 비유사성.

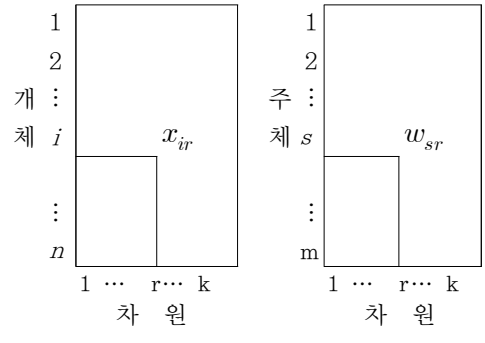


결과: k차원의 $n \times k$ 형상좌표행렬 $X_{(k)}$.

$x_{ir} \equiv$차원 r에서의 개체 i의 좌표.

주체의 $m \times k$ 가중치행렬 W.

$w_{sr} \equiv$차원 r에서 주체 s의 가중치.



$\delta_{ij}^{(s)} \equiv$ 형상공간에서 주체 s에 대한 개체 i와 j의 비유사성

〈그림 4.2〉 삼원 다차원척도법을 위한 입력자료와 결과의 도식도

특히, 식 (4.1)은 INDSCAL 모형의 차원축소된 형상공간의 비유사성이 3장에서 살펴본 유클리드거리 대신에 가중유클리드거리를 고려하고 있음을 나타낸다. 이런 이유로 INDSCAL 모형을 가중 다차원척도법(weighted MDS)이라고도 한다.

〈그림 4.1〉과 〈그림 4.2〉는 이원과 삼원 다차원척도법에 대한 입력자료와 다차원척도법 결과에서 얻어지는 형태를 비교하고 있다. 이들 그림을 입력자료와 결과의 도식도(schematic diagram)라고 한다(Young과 Lewyckyj, 1979, Chapter 1; Schiffman 외 2인, 1981, pp. 372-373; Young, 1987, pp. 122-135). 먼저 〈그림 4.1〉의 이원 다차원척도법에서 입력자료로 사용되는 비유사성행렬 D의 도형적 형태는 정사각형이며, 이는 D가 $n \times n$인 이원

정방대칭행렬임을 나타내고 있다. 그리고 k차원의 $n \times k$ 형상좌표행렬 $X_{(k)}$의 모양은 일반적으로 개체수 n이 차원수 k보다는 매우 크므로 직사각형을 이룬다. 이 형상좌표행렬로부터 형상공간에서 개체 i와 j의 비유사성 δ_{ij}를 계산할 수 있다. 다음으로 〈그림 4.2〉의 삼원 다차원척도법인 INDSCAL 모형의 입력자료는 주체×개체×개체의 삼원 행렬자료이다. 다시 말해서 각 주체에 따른 개체 간의 비유사성으로 이루어진 정사각형인 이원 정방대칭행렬이 $D^{(s)}$, $s=1,\ldots,m$이다. 그러므로 각각의 정사각형은 이원 다차원척도법의 입력자료로 볼 수 있고 이들을 각 m명의 주체에 따라 모아 두었다고 생각하면 된다. 그리고 〈그림 4.1〉에서 이원 다차원척도법의 결과는 차원축소된 k차원 형상공간에서의 개체들의 단 하나의 좌표행렬 $X_{(k)}$이지만 INDSCAL 모형의 결과는 2개의 좌표행렬로 이루어져 있다. 그 하나는 개체 형상공간에서의 좌표행렬 $X_{(k)}$와 나머지는 주체 형상공간에서의 가중치좌표행렬 W이다. 이들로부터 형상공간에서 주체 s에 대한 개체 i와 j의 비유사성 $\delta_{ij}^{(s)}$를 계산하게 된다. 특히, 개체공간을 그룹자극(group stimulus)공간이라고도 한다(Schiffman 외 2인, 1981, p. 374).

4.1.2 INDSCAL 모형의 기하적 해석

INDSCAL 모형에 바탕을 두는 삼원 다차원척도법의 기하적 해석을 이해하기 위해 다음을 생각해 보자. 삼원 다차원척도법의 적용 결과에서 차원축소된 k차원 형상공간에 개체들의 좌표행렬 $X_{(k)}=(x_{ir})$, $i=1,\ldots,n;r=1,\ldots,k$을 얻게 된다. 여기서 x_{ir}은 개체공간의 r차원에서 i번째 개체의 좌표이다. 특히, 삼원 다차원척도법에서는 식 (4.1)의 주체 s에 대하여 i번째와 j번째 개체 간의 비유사성 $\delta_{ij}^{(s)}$가 정의되기 위해서는 가중치가 필요하다. 만약에 r차원에서 s번째 주체에 대한 가중치를 w_{sr}이라고 하면 이젠 r차원에서 s번째 주체에 대한 개체공간상의 좌표는

$$y_{ir}^{(s)} \;=\; w_{sr}^{1/2}\, x_{ir} \tag{4.3}$$

에 의해서 제공된다. 식 (4.3)에 의해서 제공되는 가중 형상공간을 개인지각공간(private perceptual space)이라고 한다. 따라서 이 공간에서 주체 s에 대하여 i번째와 j번째 개체 간의 비유사성은 식 (4.3)에 의해서 유클리드거리

$$\delta_{ij}^{(s)} = \left[\sum_{r=1}^{k} (y_{ir}^{(s)} - y_{jr}^{(s)})^2 \right]^{1/2} \tag{4.4}$$

로 나타낼 수 있다. 그리고 식 (4.3)을 식 (4.4)에 대입하면 식 (4.1)을 얻게 된다. 이와 같이 식 (4.3)과 식 (4.4)는 식 (4.1)인 유클리드거리의 가중 일반화의 해석을 제공하며 대수적으로는 가중치에 의한 선형변환을 통해 INDSCAL 모형을 나타내고 있다.

다시 말해서 각 차원에 대한 각 주체의 가중치의 제곱근 $w_{sr}^{1/2}$에 의해서 각 차원이 재측정 된다. 즉 식 (4.4)에서 $w_{sr}^{1/2}$가 증가한다면 k차원에 있는 개체 간의 차이는 개체 i와 개체 j의 비유사성을 판단하는 데 큰 영향을 준다.

지금까지 설명을 기하적으로 보이기 위해 du Toit 외 2인(1986, p. 146)이 요약한 예를 살펴보자. 먼저 임의의 세 개체의 $k=2$차원의 형상좌표행렬

$$X_{(k)} = \begin{bmatrix} -0.5 & -1.6 \\ -1.3 & -1.2 \\ 1.6 & 0.8 \end{bmatrix} = (x_{ir}), \quad i=1,2,3\,;\ r=1,2$$

를 삼원 다차원척도법의 적용으로 제공되었다 하자. 그리고 두 주체에 대한 가중치행렬로

$$W = \begin{bmatrix} 0.6 & 0.4 \\ 0.8 & 0.2 \end{bmatrix} = (w_{sk}), \quad s=1,2\,;\ k=1,2$$

를 고려한다고 하자. PROC MDS에서는 이를 차원계수(dimension coefficient)라고 한다.

주체 1($s=1$)에 대한 개인지각공간의 좌표는 식 (4.3)에 의해서

$$Y^{(1)} = [y_{ir}^{(1)}] = \begin{bmatrix} w_{11}^{1/2} x_{11} & w_{12}^{1/2} x_{12} \\ w_{11}^{1/2} x_{21} & w_{12}^{1/2} x_{22} \\ w_{11}^{1/2} x_{31} & w_{12}^{1/2} x_{32} \end{bmatrix}$$

$$= \begin{bmatrix} (0.6)^{1/2}(-0.5) & (0.4)^{1/2}(-1.6) \\ (0.6)^{1/2}(-1.3) & (0.4)^{1/2}(-1.2) \\ (1.6)^{1/2}(1.6) & (0.4)^{1/2}(0.8) \end{bmatrix}$$

$$= \begin{bmatrix} -0.39 & -1.01 \\ -1.01 & -0.76 \\ 1.24 & 0.51 \end{bmatrix}$$

이다. 이와 같이 주체 2($s=2$)에 대한 개인지각공간의 좌표 $Y^{(2)}$도 계산을 할 수 있다. 이러한 좌표들의 대수적 계산과정과 기하적 해석을 위하여 [프로그램 4.1]을 이용하면 [결과 4.1]을 얻게 된다. 프로그램은 일반적인 SAS/STAT과 IML로 이루어져 있어 설명은 생략한다.

[결과 4.1]에 따르면 (a) 개체 형상공간좌표와 그림, (b) 주체별 가중치와 그림, (c) 주체 1의 개인지각공간좌표와 그림, (d) 주체 2의 개인지각공간좌표와 그림이 요약되어 있다. 이 결과는 주체별 가중치가 개체형상좌표에 기하적으로 어떻게 영향을 주는가를 나타내고 있다. 기하적으로 각 그림 (a), (c), (d)에서 개체 좌표점 o1, o2, o3를 이어 만든 삼각형을 보면 그림 (a)보다는 그림 (c)와 (d)의 것이 작다. 이는 $Y^{(1)}$와 $Y^{(2)}$의 좌표가 제공하는 그림 (c)와 (d)의 경우 (a)에 대해 (b)의 가중치가 각각 고려되어 크기가 변화된 것이다. 보다 자세히 그림 (c)를 제공하는 $Y^{(1)}$의 수치적 계산과정을 살펴보면 개체들의 형상좌표 행렬 $X_{(k)}$의 첫 번째 열의 각 원소에 가중치 $(0.6)^{1/2}=0.77$과 두 번째 열의 각 원소에는 가중치 $(0.4)^{1/2}=0.63$이 각각 곱해져 있다. 이들은 주체 1의 가중치로 그 값들이 1보다 작아 그림 (a)의 큰 삼각형 모양을 그림 (c)의 작은 삼각형 모양으로 기하적 변화를 가져왔다. 따라서 각 주체에 대한 가중치에 따라 개체 간의 비유사성이 충분히 달라질 수 있음을 알 수 있다.

[프로그램 4.1] 삼원 다차원척도법의 대수적 기하적 이해를 위한 예

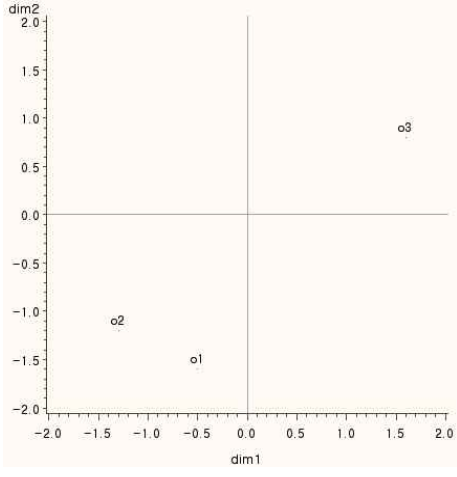
```
data config;
input id$ dim1 dim2;
cards;
o1  -0.5  -1.6
o2  -1.3  -1.2
o3   1.6   0.8
;
data weight;
input  subject$ dim1 dim2;
cards;
s1   0.6   0.4
s2   0.8   0.2
```

```
 ;
ods graphics on;
goptions reset=all;
     symbol value=point pointlabel=("#subject");
    axis1 order=(-0.0 to 1.0 by 0.1) length=4 in;
    /* Weights for Subjects Plot */
    proc gplot data=weight;
          plot dim2 * dim1/ haxis=axis1 vaxis=axis1 noframe;
proc iml;
    reset nolog;

   use config;
   read all var{id} into id;
   read all var{dim1 dim2} into X;
   n=nrow(X);
   s=nrow(W);

   use weight;
   read all var{subject} into subject;
   read all var{dim1 dim2} into W;

  W1=J(n,1,1)*sqrt(W[1,]);
  W2=J(n,1,1)*sqrt(W[2,]);
  Y1=W1#X;
  Y2=W2#X;
  print, "Objects' configuration coordinante", X[rowname=id];
  print, "Weights for Subjects", W[rowname=subject];
  print, "Subjects' configuration coordinate",Y1[rowname=id format=4.2],
                                        Y2[rowname=id format=4.2];

   vname={dim1 dim2};
   create config from X[colname=vname];
   append from X;
   close config;

   create sub1 from Y1[colname=vname];
   append from Y1;
   close sub1;

   create sub2 from Y2[colname=vname];
   append from Y2;
   close sub2;
   quit;
```

```
  data id;
    input id$@@;
    cards;
    o1 o2 o3
   ;

   data config;
         merge id config;
   data sub1;
         merge id sub1;
   data sub2;
         merge id sub2;

ods graphics on;
goptions reset=all;
symbol value=point pointlabel=("#id");
   axis1 order=(-2.0 to 2.0 by 0.5) length=4 in;
/* Objects Configuration Plot */
 proc gplot data=config;
       plot dim2 * dim1/ haxis=axis1 vaxis=axis1 href=0 vref=0 noframe;

/* Subject 1 Weighted Configuration Plot*/
proc gplot data=sub1;
      plot dim2 * dim1/ haxis=axis1 vaxis=axis1 href=0 vref=0 noframe;

/* Subject 2 Weighted Configuration Plot */
  proc gplot data=sub2;
         plot dim2 * dim1/ haxis=axis1 vaxis=axis1 href=0 vref=0 noframe;
run;
```

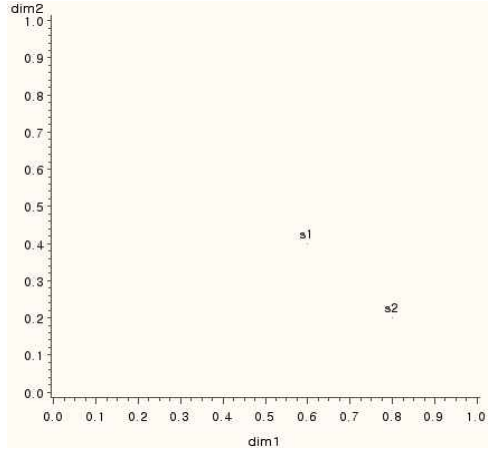
[결과 4.1] [프로그램 4.1]의 수행결과

X		
o1	-0.5	-1.6
o2	-1.3	-1.2
o3	1.6	0.8



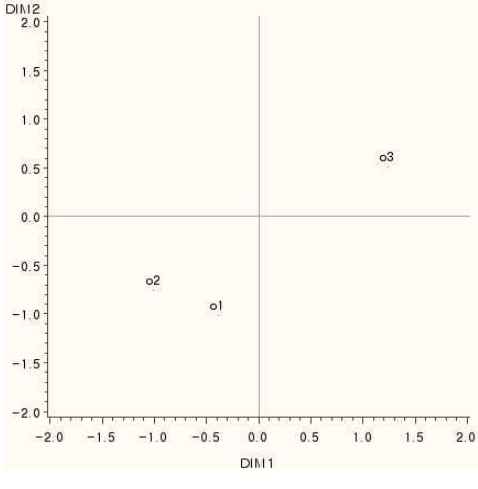
(a) 개체 형상공간좌표와 그림

W		
s1	0.6	0.4
s2	0.8	0.2



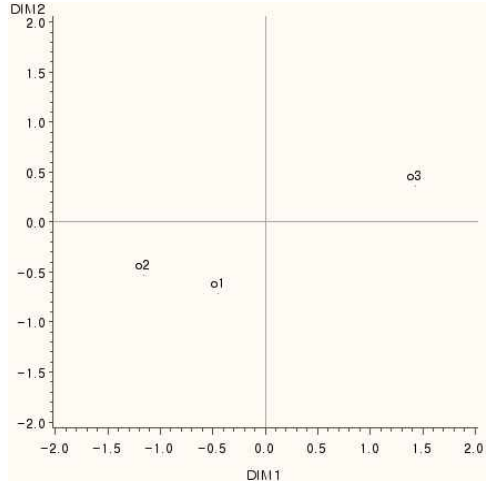
(b) 주체별 가중치와 그림

Y1		
o1	-.39	-1.0
o2	-1.0	-.76
o3	1.24	0.51



(c) 주체 1의 개인지각공간좌표와 그림

Y2		
o1	-.45	-.72
o2	-1.2	-.54
o3	1.43	0.36



(d) 주체 2의 개인지각공간좌표와 그림

사실 이 삼원 다차원척도법을 위한 INDSCAL 알고리즘을 수정하고 보완하여 Takane 외 2인(1977)은 ALSCAL을 제시하였다. 대개 INDSCAL 모형을 추정할 때 교대최소제곱(alternating least squares)법을 사용하여 이 방법의 수치해석적 절차를 ALS 알고리즘이라 한다. 그리고 이 알고리즘을 이용하는 다차원척도법을 ALSCAL이라고도 한다.

다음 절에서는 범용 SAS/PROC ALSCAL의 많은 점을 수용한 PC-윈도우용 SAS/PROC MDS 프로그램에서 제공되는 삼원 다차원척도법의 실제 예를 보이고자 한다. [부록 1.3]에서는 PROC ALSCAL과 PROC MDS를 간단히 요약 비교하고 있으니 참고하기를 바란다.

4.2 삼원 다차원척도법의 예

1.1.4절의 [사례 4] 15군데 신체 부위의 비유사성에 대한 삼원 다차원척도법의 적용에서는 이해의 관점에서 〈그림 1.4〉의 다차원척도그림을 먼저 인용하여 결과를 해석하였다. 이 절에서는 이 사례를 위한 SAS/PROC MDS [프로그램 4.2]를 제시하고 수행결과인 [결과 4.2]에서 요약된 내용들을 자세히 설명하려 한다.

4.2.1 신체 부위 자료

Young(1987, pp. 4-10)에 따르자면 Jacobowitz(1975)는 아동과 성인에게 15군데 신체 부위(cheek, face, mouth, head, ear, body, arm, elbow, hand, palm, finger, leg, knee, foot, toe)의 유사성을 서로 판단하도록 조사하였다. 이 조사에는 나이가 6세, 8세, 10세의 아동들과 대학교 2년생들로 구성되었고 각각 15명씩 배당되어 총 60명의 주체에 대해서 이루어졌다. 이는 신체에 대해 아동이 생각하는 특성과 이런 생각들이 아동이 성인이 됨에 따라 어떻게 변하는지에 대한 경향을 파악하기 위한 것이다. 이들 자료는 Young(1987, Table 1.2)이 편의상 일부를 생략해 두었고 PROC MDS를 위한 SAS Institute Inc.(1992, pp. 277-278)는 분석편의상 처음 6세 아동 15명과 대학교 2학년생 15명에 대한 자료를 제시하였지만 이것도 일부 생략된 것이다. 이미 자료에 대해서는 1.1.4절의 〈표 1.7〉에서 첫 번째 주체($s=1$)가 평가한 이들 신체 부위의 크기가 15×15인 비유사성행렬 $D^{(1)}$을 제시하여 자세히 설명하고 있다. 즉, 주체인 한 사람이 15군데 신체 부위를 각각 돌아가면서 표준부위로 두고 15점 척도로 이와 나머지 신체 부위와의 유사성을 판단한 순위-순서형(rank-order) 자료이다. 여기서는 [프로그램 4.2]와 [결과 4.2]를 중심

으로 설명하려고 한다. 추가로 완전한 자료는 Young과 Lewyckyj(1979, pp. 105-108)에 나타나 있고, 이를 [부록 2: 자료 1(파일이름: body.txt)]에 정리해 두었다. 자료에 대한 보다 자세한 설명은 Young과 Lewyckyj(1979, pp. 102-103)와 Young(1987, pp. 4-9)을 참고하기를 바란다.

4.2.2 프로그램과 결과

[프로그램 4.2]의 삼원 다차원척도법을 위한 PROC MDS 프로그램은 크게 두 부분으로 구성되어 있다.

과정 1)은 신체 부위의 비유사성행렬을 SAS자료세트 〈body〉로 제공하고 있다. 이 자료를 이용한 과정 2)는 PROC MDS 문으로 다양한 옵션에 따라 삼원 다차원척도법을 수행한다. 따라서 이 옵션과 관련하여 분석하고자 하는 내용을 설명하기로 하자. 이 자료(data=body)는 3차원(dimension=3)에서 가중유클리드모형에 바탕을 둔 INDSCAL 모형(coef=diagonal)을 적용한 순서척도(level=ordinal)로 자료의 조건은 행-조건(condition=row)으로 분석된다. 이것만으로도 [결과 4.2]의 ③ 삼원 다차원척도그림, ⑤-⑦ 차원계수그림과 ⑧ 상그림을 제공한다. 더불어 〈nophist〉에 의해서 스트레스 최소화 알고리즘의 수치적 계산과정은 출력되지 않도록 하였고, 〈pfinal〉에 의해서 ④ 최종 적합도를 제공받는다.

끝으로 삼원 다차원척도법의 특징인 주체에 대한 표시를 이미 [프로그램 4.2]의 과정 1)의 DATA 문에서 지정한 〈변수 subject의 C(아동)와 A(성인)〉로 나타내기로 한다(subject subject;). 삼원 다차원척도법과 관련하여 보다 자세한 옵션에 대한 설명은 [부록 1: SAS MDS 절차]에 잘 정리되어 있다.

[프로그램 4.2] 15군데 신체 부위자료에 대한 삼원 다차원척도법

```
1) data body;
input  cheek face mouth head ear body arm elbow hand
       palm finger leg knee foot toe;
if _n_ <= 225 then subject='C'; else subject='A';
cards;
   0   2   1   3   4  10   5   9   6   7   8  11  12  13  14
   2   0  12   1  13   3   8  10  11   9   7   4   5   6  14
   3   2   0   1   4   9   5  11   6   7   8  10  13  12  14
   2   1   3   0   4   9   5   6  11   7   8  10  12  13  14
  10   1  11   2   0   6   3   4   5  12  13   7   8  14   9
  14  12   9   6  13   0   8   7   5  10  11   1   4   2   3
  12  14  11  10  13   5   0   4   1   3   2   6   9   7   8
   5   7  14   8   6   9   1   0   2   3   4  10  11  12  13
  13  11  12  10  14   9   3   4   0   1   2   6   5   7   8
   8   6   7   9   4   5   3  10   1   0   2  12  11  13  14
  14   5  13   6   9  12   3   4   1   2   0   7   8  10  11
  14  12  13  11   9   7   4   6   5   3  10   0   8   1   2
```

```
  12  11  14  10  13   4   5   8   6   7   9   1   0   2   3
  12  14  10  13  11   9   4   5   8   6   7   2   3   0   1
  13   8   9  11  14   3   6   5   7  10  12   2   4   1   0

                                          ...

   0   1   2   4   3  14  12  10   6   5   8  13  11   7   9
   2   0   3   1   4  14   8  10   5   6  12   9  11   7  13
   2   1   0   4   3  14  12   9   5  11   7  13  10   6   8
   2   1   4   0   3  14   8  10   6  11  12   7   9   5  13
   1   3   2   4   0  14  12   9   8   7   5  13  10  11   6
  13  10  11   1  14   0   3   7   5  12   8   2   6   4   9
  13   9  11  10  12  14   0   1   3   6   5   2   4   7   8
  10  12   8  13   9  14   1   0   6  11   4   3   2   5   7
  12   9  11  10  13  14   5   4   0   1   3   8   6   2   7
   9  12  11  13  10  14   5   4   1   0   2   6   7   3   8
  10  12  11  13   9  14   6   5   1   3   0   8   7   4   2
  12  10  11   9  13  14   2   5   6   8   7   0   1   3   4
  11  12   9  13  10  14   3   2   7   8   4   1   0   5   6
  12  10  11   9  13  14   5   8   1   6   7   3   4   0   2
  10  12  11  13   9  14   8   7   4   6   2   3   5   1   0
;

2) proc mds data=body level=ordinal condition=row
 coef=diagonal dimension=3 nophist pfinal;
          subject subject;
```

[결과 4.2] [프로그램 4.2]의 수행결과:

① 형상좌표 ② 주체 가중치 ③ 삼원 다차원척도그림 ④ 최종 적합도

①

Configuration			
	Dim1	Dim2	Dim3
cheek	1.46	0.77	-0.30
face	1.40	-0.56	0.43
mouth	1.25	-0.97	-0.51
head	1.32	0.39	0.97
ear	1.22	0.03	-1.39
body	0.24	-0.18	2.24
arm	-0.70	1.23	1.08
elbow	-0.73	0.40	0.29
hand	-0.71	1.50	-0.26
palm	-0.24	1.59	-0.96
finger	-0.74	0.57	-1.35
leg	-0.69	-1.44	0.88
knee	-1.44	-0.63	0.42
foot	-0.78	-1.51	-0.36
toe	-0.87	-1.20	-1.19

②

Dimension Coefficients			
subject	1	2	3
C	1.18	1.02	0.76
C	0.94	1.05	1.01
C	0.93	1.00	1.06
C	1.08	1.10	0.79
C	0.98	1.05	0.97
C	1.28	0.85	0.80
C	0.90	1.03	1.06
C	0.95	1.04	1.01
C	1.18	1.06	0.69
C	0.95	0.99	1.05
C	1.07	1.13	0.76
C	0.98	0.94	1.08
C	1.34	0.82	0.73
C	0.95	1.00	1.05
C	1.03	1.09	0.86
A	1.30	0.82	0.80
A	1.05	0.64	1.22
A	1.28	0.85	0.81
A	0.95	0.88	1.15
A	1.34	0.75	0.80
A	1.27	0.81	0.85
A	1.50	0.37	0.78
A	1.38	0.83	0.65
A	1.44	0.56	0.78
A	1.53	0.74	0.36
A	1.19	0.95	0.84
A	0.85	0.86	1.24
A	1.29	0.62	0.97
A	1.20	0.58	1.10
A	1.24	0.62	1.04

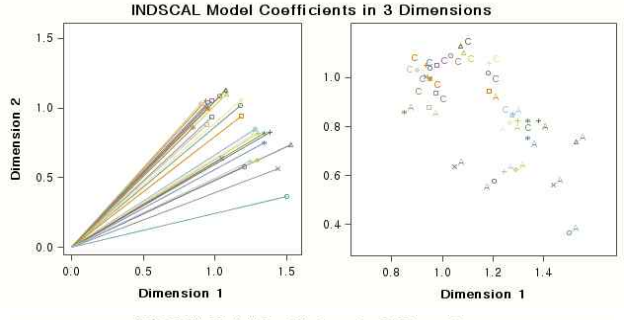
③ 〈그림 1.4〉 삼원 다차원척도그림: (a)와 (b)

④

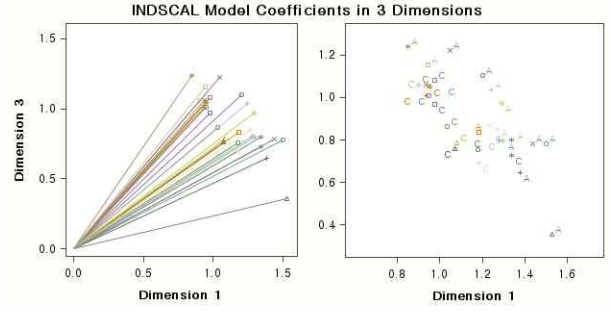
subject	Number of Nonmissing Data	Weight	Badness-of-Fit Criterion	Distance Correlation	Uncorrected Distance Correlation
C	210	0.03	0.16	0.85	0.99
C	210	0.03	0.25	0.51	0.97
C	210	0.03	0.23	0.58	0.97
C	210	0.03	0.16	0.85	0.99
C	210	0.03	0.21	0.69	0.98
C	210	0.03	0.15	0.89	0.99
C	210	0.03	0.26	0.40	0.96
C	210	0.03	0.25	0.47	0.97
C	210	0.03	0.13	0.91	0.99
C	210	0.03	0.24	0.55	0.97
C	210	0.03	0.15	0.87	0.99
C	210	0.03	0.23	0.59	0.97
C	210	0.03	0.15	0.90	0.99
C	210	0.03	0.21	0.69	0.98
C	210	0.03	0.19	0.74	0.98
A	210	0.03	0.12	0.94	0.99
A	210	0.03	0.17	0.84	0.98
A	210	0.03	0.12	0.93	0.99
A	210	0.03	0.21	0.68	0.98
A	210	0.03	0.14	0.91	0.99
A	210	0.03	0.16	0.87	0.99
A	210	0.03	0.09	0.98	1.00
A	210	0.03	0.11	0.95	0.99
A	210	0.03	0.11	0.96	0.99
A	210	0.03	0.13	0.95	0.99
A	210	0.03	0.13	0.90	0.99
A	210	0.03	0.21	0.72	0.98
A	210	0.03	0.12	0.93	0.99
A	210	0.03	0.10	0.95	0.99
A	210	0.03	0.11	0.94	0.99
All	6300	1.00	0.17	0.84	0.98

[결과 .2] (연속) ⑤-⑦ 차원계수그림 ⑧ 상그림

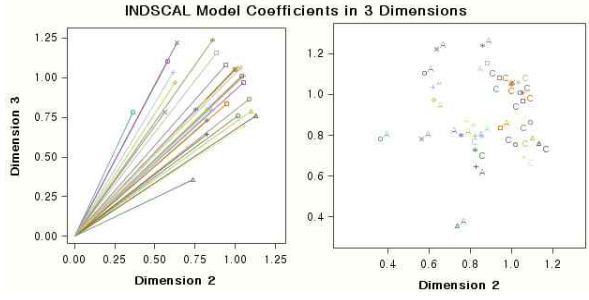
⑤



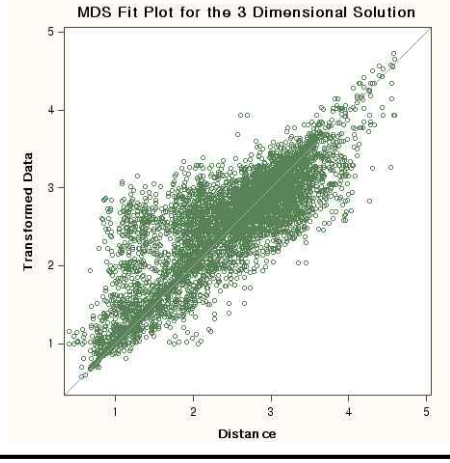
⑥



⑦



⑧



4.2.3 결과의 해석

[결과 4.2]의 ①은 15군데 신체 부위에 대한 3차원 형상공간의 좌표이고 ②는 각 주체에 대한 차원별 가중치이다. 이들의 역할은 이미 4.1.2절에서 예를 들어 기하적 해석과 함께 보였다. 결과 ③의 〈그림 1.4〉 삼원 다차원척도그림: (a)와 (b)의 해석은 1.1.4절에 자세히 설명되어 있으니 참고 바란다. 결과 ④에서는 각 아동(C)과 성인(A) 그룹별 주체 15명씩 총 30명에 대한 스트레스(Badness-of-Fit Criterion)가 1.3.2절 식 (1.7)의 크루스칼의 $ST1$ 공식에 의해 제공되었고 이들의 평균 스트레스값

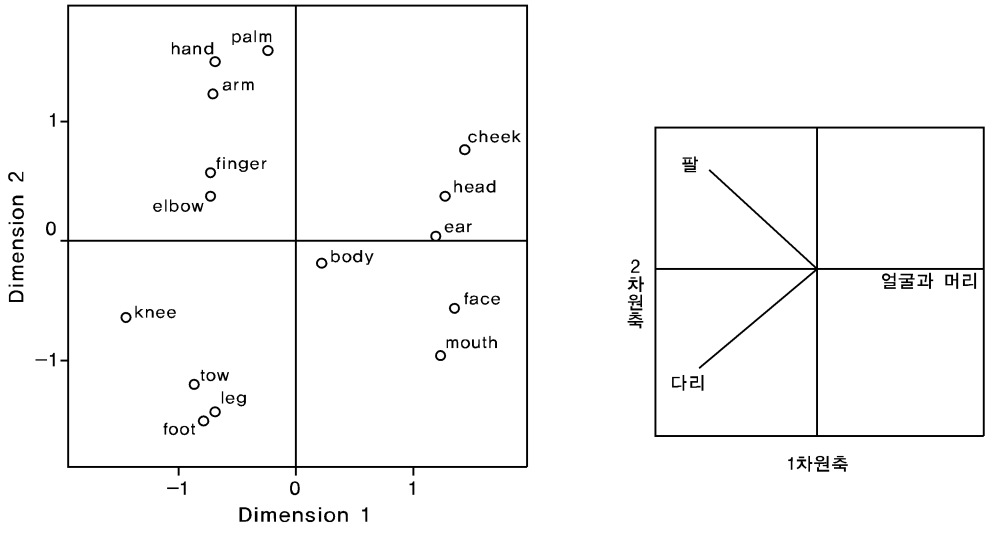
$$\frac{0.16 + 0.25 + \cdots + 0.11}{30} = 0.1666 \simeq 0.17$$

이 최종 적합도로 맨 아래 〈-ALL-〉에 나타나 있다. 이 값은 차원수 결정을 위한 크루스

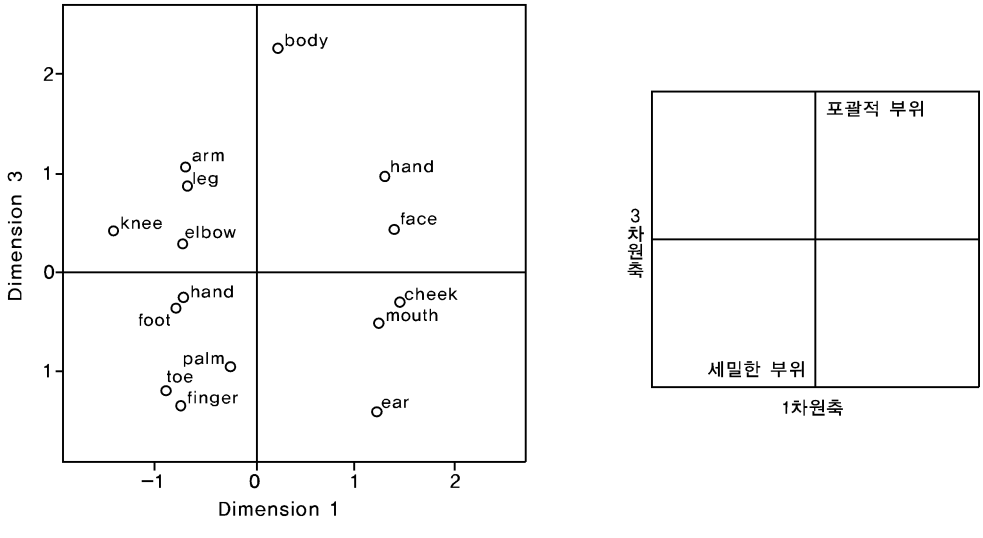
칼의 판별기준에 따르면 보통(0.10)과 나쁨(0.20) 사이에 있어 2차원의 삼원 다차원척도법의 결과가 자료를 어느 정도 요약해 준다고 여겨진다.

결과 ⑤-⑦ 차원계수그림은 차원별로 조합되어 주어져 있는데 이들의 영향으로 형상공간좌표의 기하적 변화를 4.1.2절에서 자세히 설명한 바 있다. 실제로 이들 차원계수그림은 사실 결과 ②의 각 주체별로 주어진 가중치로 이루어진 차원계수좌표로 그려진 것이다. 참고로 이들과 관련된 삼원 다차원척도그림에 대한 그룹별 해석을 위하여 〈그림 1.4〉를 〈그림 4.3〉으로 여기서 다시 인용하기로 하자. 먼저 결과 ⑤-⑦을 보면 2차원축은 아동(C)이 성인(A)과 다름을 보여 주고 있다. 아동들은 각 차원에서 가중치가 거의 1.0으로 대동소이하나 성인들은 아동보다 다양하다. 특히 성인들의 가중치는 2차원축에 대해서 1.0보다 작으며 실제 이들의 평균은 약 0.7이다. 따라서 이런 사실로부터 〈그림 4.3〉의 (a) 1차원과 2차원의 삼원 다차원척도그림의 2차원축(팔과 다리부위를 분할하는 축)에서 본다면 성인들은 아동들이 생각하는 것보다 팔과 다리부위가 더욱 더 비슷하다고 생각한다. 그리고 대다수 성인들이 결과 ⑤의 1차원축에 높은 가중치를 받고 있는데, 이를 〈그림 4.3〉-(a) 삼원 다차원척도그림의 1차원축(팔과 다리부위와 머리부위로 분할하는 축)에서 본다면 성인들은 아동들이 생각하는 것보다 머리부위와 팔과 다리부위가 더욱 더 비슷하지 않다고 생각한다는 점을 알 수 있다. 결과 ⑥의 3차원축에서는 아동과 성인의 가중치가 다양하므로 〈그림 4.3〉-(b) 1차원과 3차원에 의한 삼원 다차원척도그림의 3차원축(신체의 포괄적 부위와 세밀한 부위를 분할하는 축)의 관점에서 아동과 성인 모두가 신체 부위에 대해서는 상당히 다양한 판단을 내림을 알 수 있다. 지금까지의 분석의 내용은 SAS Institute, Inc.(1992, pp. 277-285; 2008, pp. 3726-3735)를 참고하였다.

끝으로 결과 ⑧ 상그림은 삼원 다차원척도법의 적합성을 위한 Guttman의 상그림이다. 이 그림에서 원점을 지나는 직선상에 좌표점들이 많이 놓여 있어 삼원 다차원척도법을 위한 가중유클리드모형이 적합함을 알 수 있다. 이 Guttman의 상그림에 대해서는 3.2.2절에서 자세히 소개하였다. 특히, SAS/PROC MDS 9.2판에서 결과 ⑧ 상그림은 자동으로 주어지는데, 이 버전 이전에는 옵션 〈out=out〉과 〈outres=res〉에 의해서 다차원척도그림을 위한 형상좌표에 관련된 추정결과와 모형의 적합성을 보여 주는 값을 저장한 자료세트를 만들어 상그림과 세퍼드그림을 얻어내었다(최용석, 1995, 4장).



(a) 1차원과 2차원



(b) 1차원과 3차원

〈그림 4.3〉 (〈그림 1.4〉의 재인용) 15군데 신체 부위에 대한 삼원 다차원척도그림과 해석도 : (a) 1차원과 2차원 (b) 1차원과 3차원

연습문제

4.1 다음은 두 명의 학생들이 여덟 나라의 비유사성을 판단한 자료이다(Davison, 1992, pp. 147-149).

학생 1	Angola	.
	Argentina	1.41 .
	Australia	1.00 1.00 .
	China	1.00 1.73 1.41 .
	Cuba	1.41 1.41 1.73 1.00 .
	Japan	1.41 1.41 1.00 1.00 1.41 .
	United States	1.73 1.00 1.41 1.41 1.00 1.00 .
	Zimbabwe	0.71 1.41 1.00 1.00 1.41 1.41 1.73 .
학생 2	Angola	.
	Argentina	1.00 .
	Australia	2.00 2.00 .
	China	3.00 3.00 2.00 .
	Cuba	1.00 1.00 3.00 1.00 .
	Japan	2.00 2.00 1.00 2.00 1.00 .
	United States	3.00 3.00 2.00 3.00 2.00 2.00 .
	Zimbabwe	1.00 1.00 3.00 1.00 3.00 3.00 1.00 .

1) 적절한 차원수 결정을 위하여 1.3.3절에서 언급한 스트레스-차원수 그림을 제공하라.

2) 삼원 다차원척도법을 실시하고 모형의 적합성을 제시하는 상그림을 해석하라.

3) 삼원 다차원척도그림을 제공하고 해석하라.

4) 차원계수그림과 더불어 다차원척도그림을 추가 해석하라.

4.2 다음은 두 사람이 네 가지 식용작물(감자, 시금치, 상치, 튜나)에 대한 비유사성의 순위를 매겨 얻은 자료(Schiffman 외 2인, 1981, p. 86; Young, 1987, p. 137)로 원래 다섯 사람에 대한 자료이다. 이 자료는 아래쪽 삼각형이 첫 번째 사람, 위쪽 삼각형이 두 번째 사람에 대한 것이다. 삼원 다차원척도법을 실시하고 해석하라.

감자	.	1	3	6
시금치	4	.	2	5
상치	2	1	.	4
튜나	3	6	5	.

4.3 다음은 건강, 과학, 기술, 상업, 사무, 중개, 사회, 예술 여덟 가지 관심 영역에 대한 네 그룹 각각의 심리적 시험 점수에 대한 상관계수를 구한 것이다(du Toit 외 2인, 1984, p. 148). 네 그룹은 유색 여성, 아시아 여성, 남아프리카 공용 네덜란드어를 사용하는 백인 여성, 영어를 사용하는 백인 여성으로 이루어져 있다. [부록 2: 자료 8(파일이름: field.txt)]를 참고 바란다.

1) 스트레스-차원수 그림을 통하여 적절한 차원수를 결정하라.
2) 네 그룹 각각에 대한 이원 다차원척도법을 실시하고 관심 영역 간의 유사성을 비교하라.
3) 삼원 다차원척도법을 실시하고 상그림을 통하여 모형의 적합성을 논의하라.
4) 삼원 다차원척도그림을 제공하고 해석하라.

유색 여성	건강	1.00
	과학	0.71 1.00
	기술	0.72 0.67 1.00
	상업	0.60 0.55 0.88 1.00.
	사무	0.55 0.60 0.55 0.55 1.00
	중개	0.63 0.41 0.69 0.71 0.76 1.00
	사회	0.67 0.54 0.61 0.56 0.66 0.76 1.00
	예술	0.63 0.58 0.76 0.76 0.49 0.64 0.59 1.00
아시아 여성	건강	1.00
	과학	0.80 1.00
	기술	0.54 0.72 1.00
	상업	0.34 0.52 0.84 1.00.
	사무	0.35 0.43 0.66 0.69 1.00
	중개	0.33 0.45 0.62 0.67 0.78 1.00
	사회	0.52 0.65 0.60 0.50 0.54 0.66 1.00
	예술	0.32 0.53 0.79 0.76 0.64 0.72 0.59 1.00
남아프리카 공용 네덜란드어 사용 백인 여성	건강	1.00
	과학	0.77 1.00
	기술	0.55 0.74 1.00
	상업	0.47 0.61 0.85 1.00.
	사무	0.23 0.23 0.49 0.53 1.00
	중개	0.37 0.36 0.62 0.60 0.74 1.00
	사회	0.59 0.52 0.55 0.49 0.47 0.74 1.00
	예술	0.48 0.59 0.72 0.69 0.43 0.64 0.63 1.00
영어를 사용 백인 여성	건강	1.00
	과학	0.72 1.00
	기술	0.53 0.69 1.00
	상업	0.45 0.51 0.77 1.00.
	사무	0.17 0.16 0.45 0.40 1.00
	중개	0.31 0.25 0.49 0.38 0.77 1.00
	사회	0.54 0.41 0.47 0.37 0.44 0.67 1.00
	예술	0.38 0.50 0.67 0.55 0.36 0.49 0.55 1.00

5장

최대우도 다차원척도법

다차원척도법에서 Ramsay(1977, 1980, 1982)가 제안한 최대우도 다차원척도법(maximum likelihood MDS)에 대한 모형을 램지모형이라고 부르기도 하는데, 비유사성의 로그-정규분포를 가정하고 최대우도추정법이 사용된다. 최대우도 다차원척도법도 4장의 삼원 다차원척도법과 같이 주체에 대한 개체 간의 삼원 형태의 비유사성행렬의 분석에 활용되며, 사례를 통하여 서로 비교하고 그 차이점을 이해하려고 한다. 그리고 최대우도 다차원척도법을 위한 SAS/PROC MDS 프로그램과 그 결과를 제시하여 활용도를 높이고자 한다.

5.1 최대우도 다차원척도법을 위한 모형

최대우도 다차원척도법은 Ramsay(1977, 1980, 1982)가 제안하였다. 이를 램지모형이라고 한다. Young(1987, pp.36-39)은 다차원척도법에서 최대우도기법이 다차원척도법을 자료분석의 기술적이며 탐색적 도구에서 추론적 도구로 변화 발전시켰다고 평가한다. 그는 덧붙여 적절한 가정 아래 최대우도 다차원척도법은 차원수, MDS 모형 그리고 오차모형을 적절하게 결정하는 유의성 검정도 제공한다고 하였다.

5.1.1 램지모형의 소개

Ramsay(1977, 1980, 1982)는 최대우도 다차원척도법을 위한 그의 모형에 대해서 설명하고 있는데, 이 절에서 요약해 보자.

일반적으로 삼원 다차원척도법에서와 같이 m명의 주체에 대해서 고려한 삼원행렬 형태인 비유사성행렬

$$D^{(s)} = (d_{ij}^{(s)}), \quad i,j = 1,\ldots,n;\ s = 1,\ldots,m$$

에서 $d_{ij}^{(s)}$는 s번째 주체에 대해 i번째와 j번째 개체 사이의 가중유클리드거리에 의한 비유사성을 나타낸다고 하자. 최대우도 다차원척도법에서도 궁극적으로 구하고자 하는 $k\ (\le p)$ 차원 형상공간에서 크기가 $n \times k$인 형상좌표를 $X_{(k)} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^t$라고 하면 $\boldsymbol{x_i} = (x_{i1},\ldots,x_{ik})^t$는 크기가 $k \times 1$인 벡터로 k차원에서 개체 $i(=1,\ldots,n)$의 좌표점이 된다. $s(=1,\ldots,m)$번째 주체에 대한 r번째와 v번째 차원에 대한 가중치를 $w_{rv}^{(s)}$라고 하면

$$W_s = (w_{rv}^{(s)}), \quad r,v = 1,\ldots,k;\ s = 1,\ldots,m \tag{5.1}$$

는 크기가 $k \times k$인 가중치행렬이 된다. 따라서 가중치를 고려한 개인지각공간에서 i번째와 j번째 개체 간의 비유사성을

$$\delta_{ij}^{(s)} = \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)^t W_s (\boldsymbol{x}_i - \boldsymbol{x}_j)\right]^{1/2} \tag{5.2}$$

라 하면 최대우도 다차원척도법을 위한 모형은 식 (4.2)와 같이

$$d_{ij}^{(s)} \;=\; f_s(\delta_{ij}^{(s)}) + \epsilon_{ij}$$

를 만족한다. 여기서 $f_s(\cdot)$는 일반적으로 단조함수이고 ϵ_{ij}는 왜곡오차이다.

Ramsay(1982)는 특히 식 (5.1)에서 가중치행렬이 $W_s = I$인 단위행렬이라면 항등계량형(identity metric)모형이라고 하였고 $W_s = diag(w_{11}^{(s)},...,w_{kk}^{(s)})$인 경우는 대각계량형(diagonal metric)모형이라고 하였다. 만약 대각원소를 $w_{11}^{(s)} = w_{s1}, w_{22}^{(s)} = w_{s2},..., w_{kk}^{(s)} = = w_{sk}$로 두면 4장의 삼원 다차원척도법을 위한 INDSCAL 모형에서 식 (4.1)의 비유사성과 일치함을 알 수 있다.

그는 최대우도 다차원척도법의 모형에서 단조함수 $f_s(\cdot)$를 추정하기 위해 비유사성 $d_{ij}^{(s)}$의 로그-정규분포(lognormal distribution)

$$\log d_{ij}^{(s)} \;\sim\; \mathrm{N}(\log\delta_{ij}^{(s)},\; \sigma_{ij(s)}^2\;) \tag{5.3}$$

가정 아래 최대우도추정(MLE) 과정을 유도하고 있다. Everitt와 Dunn(1991, pp. 91-95)은 이런 과정의 주된 부분을 요약하고 있는데 이를 단계별로 구성하여 [알고리즘 4: 램지 최대우도추정법]을 제시하고 있다. 따라서 램지 최대우도 다차원척도법은 [알고리즘 4]의 [2 단계]에서 가정한 모형의 오차분산 $\sigma_{ij(s)}^2 = \sigma_{(s)}^2\gamma_{ij}^2$를 추정하고 저차원 형상공간의 좌표를 얻는데 최대우도추정법을 적용하는 최대우도추정과정을 따르고 있다. 여기서 $\sigma_{(s)}^2$는 주체특수분산(subject specific variance)이고 γ_{ij}^2은 짝특수분산인자(pair specific component)이다.

[알고리즘 4] 램지 최대우도추정법
1단계: 비유사성 $d_{ij}^{(s)}$를 구하고 $\log d_{ij}^{(s)}$를 계산한다.
2단계: 로그-정규분포 $\log d_{ij}^{(s)} \;\sim\; \mathrm{N}(\log\delta_{ij}^{(s)},\; \sigma_{ij(s)}^2)$를 가정하고 $f_s(\log d_{ij}^{(s)})$와

$\log\delta_{ij}^{(s)}$의 관계는 다음과 같다.

$$f_s(\log d_{ij}^{(s)}) = \log\delta_{ij}^{(s)} + \epsilon_{ij}^{(s)} \tag{5.4}$$

여기서 $f_s(\cdot)$는 단조함수이고 오차 $\epsilon_{ij}^{(s)}$는 서로 독립이면서 $\mathrm{N}(0, \sigma_{ij(s)}^2)$를 따르며 $\sigma_{ij(s)}^2 = \sigma_{(s)}^2\gamma_{ij}^2$를 고려한다.

3단계: 식 (5.4)를 만족하는 $f_s(\log d_{ij}^{(s)})$에 대해 $\log\delta_{ij}^{(s)}$가 최적적합을 이루고 동시에 분산 $Var(\epsilon_{ij}^{(s)}) = \sigma_{ij(s)}^2$를 추정하여 거리와 좌표를 구하고 저차원 형상공간좌표를 얻는다.

Ramsay(1982)는 이를 각각 〈within-subject variance〉와 〈within-pair variance factor〉라고 하였다. γ_{ij}^2는 다양한 방법으로 분해되거나 1이 될 수도 있다. 그중에서 Ramsay(1982)는 다음의 짝특수분산 $\gamma_{ij}^2 = (a_i^2 + a_j^2)/2$을 고려하고 있다. 여기서 계수 a_i, $i = 1,...,n$는 $\Sigma_{i=1}^n a_i^2 = n$을 만족하며 주체 s에 대한 각 개체의 상대적 인지도를 나타낸다. 즉 개체 i와 개체 j의 상대적 인지도가 주체 s에 대해 크다면 잔차의 분산을 크게 하게 된다. Ramsay(1982)는 이 분산을 추정하는 데 세 가지 접근방법(conditional moment estimate, MINQUE theory, maximum likelihood)을 제시하고 있다.

5.1.2 등분산 가정에서 최대우도추정

5.1.1절의 최대우도추정법인 [알고리즘 4]의 [2단계]에서 Ramsay(1977)는 편의성에 따라 오차의 등분산을 가정하여 자세히 요약하고 있으며 이를 간단히 정리해 보기로 하자.

크기가 $k \times 1$인 벡터 $\boldsymbol{x}_i = (x_{i1},..., x_{ik})^t$를 k 차원 형상공간에서 개체 $i(=1,...,n)$의 좌표점이라면 이들에 의해 개체 간의 비유사성 $\delta_{ij}^{(s)}$를 식 (5.2)에 의해서 정의된다고 하자. 그러면 실제 개체 간의 비유사성 $d_{ij}^{(s)}$의 분포는 독립으로 $f(d_{ij}^{(s)} \mid \delta_{ij}^{(s)}, \sigma^2)$를 따른다고 가정하게 된다. 독립성이란 가정에 의해서 로그우도함수(log likelihood function)는

$$\log L = \sum_i \sum_j \sum_s \log f(d_{ij}^{(s)} \mid \delta_{ij}^{(s)}, \sigma^2) \tag{5.5}$$

이다. 따라서 최대우도추정은 개체 i의 좌표점 $\boldsymbol{x}_i = (x_{i1}, ..., x_{ik})^t$가 k개 좌표 $x_{i1}, ..., x_{ik}$로 이루어져 있어 전체 n개 개체에 대하여 nk개의 좌표점과 분산 σ^2에 대해 식 (5.5)의 로그우도함수 $\log L$을 최대화하는 것이다. 여기서 nk개의 좌표점은 $k\ (\leq p)$차원 형상공간에서 크기가 $n \times k$인 형상좌표 $X_{(k)} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^t$를 이룬다. 식 (5.5)에서 대개 확률밀도함수 $f(\cdot)$로는 로그정규분포

$$f(\log d_{ij}^{(s)} \mid \delta_{ij}^{(s)}, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} exp\left[-\frac{1}{2\sigma^2} \left(\log d_{ij}^{(s)} - \log \delta_{ij}^{(s)} \right)^2 \right]$$

를 취한다. 이는 $\log d_{ij}^{(s)} \sim \mathrm{N}(\log \delta_{ij}^{(s)},\ \sigma^2)$이며 식 (5.3)에서 $\sigma_{ij(s)}^2 = \sigma^2$인 경우와 동일하다.

식 (5.5)로부터 로그우도함수는 다음과 같다.

$$\log L = -n^2 m \log \sqrt{2\pi} - n^2 m \log \sigma - \frac{1}{2\sigma^2} \mathrm{S}$$

여기서 $\mathrm{S} = \sum_i \sum_j \sum_s \log^2 (d_{ij}^{(s)} / \delta_{ij}^{(s)})$이다. 따라서 σ^2에 대한 최대우도추정치는 $\widehat{\sigma^2} = \mathrm{S}/n^2 m$이다.

이외에도 Ramsay(1977)는 근사적 불편추정치와 베이즈추정치를 제시하고 있다. 특히 식 (5.1)에서 $W_s = diag(w_{11,}^{(s)} ..., w_{kk}^{(s)}) = diag(w_{s1}, ..., w_{sk})$인 대각계량형모형인 경우 $\Sigma_{s=1}^m w_{sr}^2 = \mathrm{N}$, $r = 1, ..., k$라는 제약을 두자. 그러면 형상좌표행렬 $X_{(k)}$와 가중치행렬 W_s의 최대우도추정치는

$$\mathrm{Q} = \log L - \frac{1}{2} \sum_{r=1}^{k} \mu_r \left(\sum_{s=1}^{m} w_{sr}^2 - \mathrm{N} \right)$$

를 최대화 하도록 찾으면 되고 여기서 μ_r, $r=1,\ldots,k$를 라그랑지승수라고 한다. 그리고 그 결과는 다음과 같다.

$$x_{pq}=\sum_j c_{pjq}\,x_{jq},$$

$$w_{sq}=\left[\sum_i\sum_j u_{ij(s)}(x_{iq}-x_{jq})^2\right]/(2\mu_q).$$

여기서

$$c_{pjq}=\sum_s u_{pj(s)}w_{sq}/\sum_i\sum_s u_{pi(s)}w_{sq},$$

$$\mu_q=\left[\mathrm{N}^{-1}\sum_s\left\{\sum_i\sum_j u_{ij(s)}(x_{iq}-x_{jq})^2\right\}^2\right]^{1/2},$$

$$u_{pj(s)}=\sigma^2(\delta_{pj}^{(s)})^{-1}\frac{\partial}{\partial\delta_{pj}^{(s)}}\left[\log f(d_{pj}^{(s)})+\log f(d_{jp}^{(s)})\right]$$

이다. Ramsay(1977, 1980, 1982)의 논문들은 최대우도 다차원척도법의 여러 가지 면을 잘 다루고 있으니 관심 있는 독자는 참고하기를 바란다.

램지 최대우도 다차원척도법은 범용 SAS에서 PROC MLSCALE로 1986년에 채택되었다. 이와는 별도로 프로그램 MULTISCALE도 자체 개발하였다. 여기에 대한 자세한 설명을 Schiffman 외 2인(1981, pp. 389-405)에 램지가 직접하고 있다. 비록 PC-윈도우즈용 PROC MDS가 PROC MLSCALE의 몇 가지 부분을 받아들였지만 똑같은 결과를 제공하지는 못한다. 여기에 대해서는 [부록 1.4]에서 비교하고 있으니 참고하기를 바란다. 근본적으로 PROC MDS는 추정방법으로 최소제곱법을 사용한다. 그러나 최소제곱법도 대개 오차항이 서로 독립이고 등분산인 표준정규분포를 따른다면 최대우도법과 같은 결과를 제공함은 잘 알려져 있다.

5.2 최대우도 다차원척도법의 예

이미 1.1.5절의 [사례 5]에서는 10명의 사람이 15가지 레크리에이션의 비유사성을 평가한 자료에 대하여 최대우도 다차원척도법을 적용하였고 〈그림 1.5〉의 최대우도 다차원척도 그림을 제공하여 해석을 하였다. 이 절에서는 이 사례를 위한 [프로그램 5.1]을 제공하고 [결과 5.1]에서 그 수행 결과를 요약하고 해석하려고 한다.

5.2.1 레크리에이션 자료

Ramsay(1983)는 10명에게 15가지 레크리에이션(concert, museum, theatre, movie, watch TV, conference, reading, watch hockey, ballet, political debate, fashion show, documentary film, exhibition, window shopping, restaurant)의 비유사성을 판단하기 위해 조사하였다. 이 조사는 레크리에이션을 각각 둘씩 짝지은 105가지 경우에 대해 다음과 같이 비유사성의 정도를 25점 척도로 표시한 설문지로부터 얻어진 것이다.

예) Museum과 Hockey		
매우 유사하다(0점) (very similar)	_____	매우 다르다(24점) (very different)

[부록 2: 자료 2(파일이름: recreation.txt)]에 크기가 15×15인 비유사성행렬 10개가 정리되어 있다. 이것은 설문 응답자 10명의 주체에 대하여 삼원행렬자료 $D^{(s)}$, $s = 1, \ldots, 10$에 해당한다. 이 중에서 첫 번째 한 사람의 평가로 이루어진 비유사성행렬 $D^{(1)}$을 1.1.5절의 〈표 1.8〉에서 참고로 제공하였고 이 행렬의 구성에 대해서 자세히 설명하고 있으니 참고하기를 바란다.

다음으로 이 자료에 대한 최대우도 다차원척도법을 위한 프로그램을 5.2.2절에서 본격적으로 살펴보자.

5.2.2 프로그램과 결과

[프로그램 5.1]은 최대우도 다차원척도법을 5.2.1절에서 설명한 자료에 적용하기 위한 SAS/PROC MDS 프로그램이다. 이는 크게 두 과정으로 구성되어 있는데, 과정 1)은 램지

의 레크리에이션 자료를 10명의 각 사람별로(if _n_ 〈16 then subject=1; … else subject=10;) 비유사성자료를 읽어 들여 SAS 자료세트 〈recreat〉 이름으로 저장하게 한다. 과정 2)는 이 자료(data=recreat)를 이용하여 가중유클리드거리(coef=diagonal)를 이용한 최대우도 다차원척도법(fit=log level=loginterval)으로 입력자료의 조건과 행렬형태는 디폴트인 〈condition=matrix shape=triangular〉가 사용되었다. 추가적으로 옵션 〈pfinal nophist〉는 스트레스 최소화 알고리즘의 최종 수치적 결과만 제공하라는 것이다.

[프로그램 5.1] 레크리에이션 자료에 대한 최대우도 다차원척도법

```
1) data recreat;
        input (concert museum theatre movie watch_tv conferen
                reading hockey ballet politic fashion document
                exhibit shopping restaura ) (3.0) @45;
        if _n_ < 16 then subject= 1;
   else if _n_ < 31 then subject= 2;
   else if _n_ < 46 then subject= 3;
   else if _n_ < 61 then subject= 4;
   else if _n_ < 76 then subject= 5;
   else if _n_ < 91 then subject= 6;
   else if _n_ < 106 then subject= 7;
   else if _n_ < 121 then subject= 8;
   else if _n_ < 136 then subject= 9;
   else subject=10;
   cards;
     0
     8   0
    18  21   0
     5  22  16   0
     8  24  22  11   0
    11  13  19  18  22   0
    19  12  15   9  21  14   0
     6   6  15  19  21  13  23   0
    24  18  21  15   9  19  16  23   0
    21  21  23  23  23  13  12  11  23   0
    22  20  24  21  23  10  13  12   7   7   0
    10   7  21  11   9  10   6   8  14  22  22   0
    23  21  18   9   8  10   7  14  22   9  11   8   0
     5  14  19  16  23  22  19  20   8  17  23  21  20   0
    14  23  11   8  21  22   7  10  23  19  21   9   7  22   0
                 …
```

```
      0
     22   0
     25   9   0
      3  22  25   0
     14  25  16   7   0
     18  22   7  20  24   0
     22   5  11  14  22   5   0
      9   6  11  22  22   4  19   0
     25  21  18   4  13  18  14  22   0
     22  21  23  16  19  20  21  20  12   0
     19  22  23  21  23   7  21  15   7  18   0
      4  19  23  21  12  22   7   6  19  21  12   0
      9   6  16  12  18  22  22  13  21  20  21   6   0
     21   6  22  22   5  23   4  22  17  19  23  20  17   0
     16  22  17   4   7   5  16  16  21   6  21   5  23  17   0
    ;
    run;

2)  ods graphics on;
    proc mds data=recreat fit=log level=loginterval coef=diagonal
                    pfinal nophist;
          subject subject;
```

프로그램의 수행결과는 [결과 5.1]의 ① 형상좌표, ② 주체 가중치, ③ 변환함수의 기울기와 멱승, ④ 최대우도 다차원척도그림, ⑤ 최종 적합도, ⑥ 차원계수그림, ⑦ 상그림을 제공한다. 삼원 다차원척도법과 같이 최대우도 다차원척도법도 각 주체별 비유사성행렬을 분석하는 것이므로 실제 분석에 SUBJECT 문(subject subject;)을 사용하였다. 최대우도 다차원척도법과 관련된 프로그램의 문법과 옵션에 대해서는 [부록 1: SAS MDS 절차]를 참고하기를 바란다.

[결과 5.1] [프로그램 5.1]의 수행결과: ① 형상좌표, ② 주체 가중치, ③ 변환함수의 기울기와 멱승, ④ 최대우도 다차원척도그림, ⑤ 최종 적합도, ⑥ 차원계수그림

①

Configuration		
	Dim1	Dim2
concert	0.01	1.06
museum	0.19	1.51
theatre	1.38	-0.14
movie	1.28	-0.46
watch_tv	0.77	-1.92
conferen	0.19	0.80
reading	0.19	-0.13
hockey	0.01	1.37
ballet	-1.56	-1.39
politic	-1.28	-0.34
fashion	-1.56	-1.17
document	0.01	0.52
exhibit	0.59	-0.55
shopping	-1.49	1.06
restaura	1.28	-0.20

②

Dimension Coefficients		
subject	1	2
1	1.01	0.99
2	0.96	1.04
3	0.70	1.23
4	1.41	0.00
5	1.20	0.76
6	0.78	1.18
7	0.79	1.18
8	0.98	1.02
9	1.04	0.96
10	1.41	0.00

③

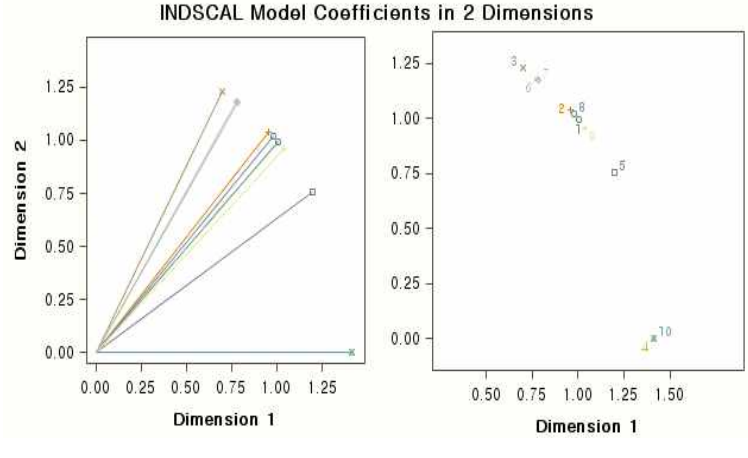
subject	Slope	Power
1	12.29	0.35
2	9.89	0.50
3	9.98	0.45
4	15.36	0.11
5	11.43	0.36
6	11.74	0.28
7	11.22	0.50
8	11.90	0.24
9	8.50	0.42
10	14.62	0.07

④ 〈그림 1.5〉 최대우도 다차원척도그림

⑤

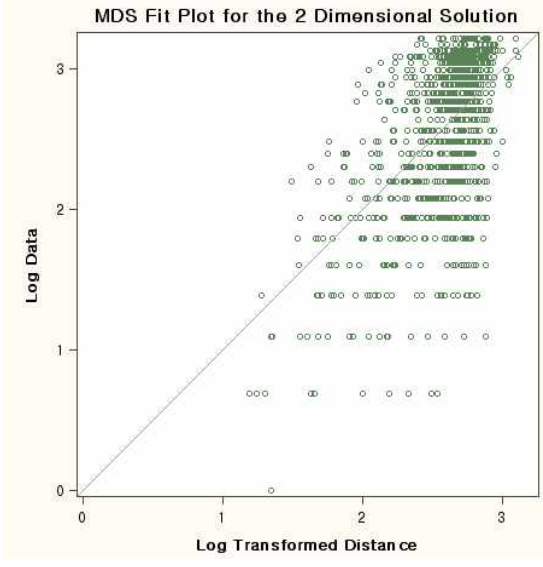
subject	Number of Nonmissing Data	Weight	Badness-of-Fit Criterion	Distance Correlation	Uncorrected Distance Correlation	Fit Correlation	Uncorrected Fit Correlation
1	105	0.10	0.84	0.56	0.96	0.55	0.99
2	105	0.10	0.78	0.54	0.95	0.62	0.99
3	105	0.10	0.85	0.48	0.92	0.52	0.98
4	105	0.10	0.76	0.54	0.95	0.66	0.99
5	105	0.10	0.88	0.42	0.93	0.47	0.98
6	105	0.10	0.91	0.38	0.94	0.41	0.99
7	105	0.10	0.75	0.59	0.95	0.66	0.99
8	105	0.10	0.94	0.31	0.94	0.35	0.99
9	105	0.10	0.88	0.45	0.92	0.48	0.97
10	105	0.10	0.92	0.35	0.93	0.39	0.98
All	1050	1.00	0.85	0.49	0.94	0.54	0.99

⑥



[결과 5.1] (연속) ⑦ 상그림

⑦



5.2.3 결과의 해석

[결과 5.1]을 살펴보면 결과 ①은 15가지 레크리에이션에 대한 2차원 형상좌표이며 10명의 각 사람에 대한 차원별 가중치가 결과 ②에 주어져 있다.

결과 ③은 각 사람별 멱승(power)변환

$$f_s(\log d_{ij}^{(s)}) = p_s \log d_{ij}^{(s)} + v_s$$

에서 기울기(slope) v_s와 멱승(power) p_s의 값을 나타낸다. 단조함수의 변환의 유형은 크게 척도(scale)와 멱승이 있다. 척도변환은 다음과 같다.

$$f_s(\log d_{ij}^{(s)}) = \log d_{ij}^{(s)} + v_s.$$

이는 [알고리즘 4]에서 $f_s(\log d_{ij}^{(s)})$에 대해 $\log \delta_{ij}^{(s)}$가 최적적합되도록 단조변환함수 $f_s(\cdot)$를 찾게 해 준다.

결과 ④의 최대우도 다차원척도그림은 1.1.5절의 [사례 5]에서 〈그림 1.5〉로 해석도와 함께 축에 대한 해석과 더불어 이미 제시된 것으로 편의상 다음 절인 5.3절에서 삼원 다차원척도그림과의 비교를 위해 재인용하고 있다.

결과 ⑤는 각 사람에 대한 스트레스 Badness-of-Fit Criterion으로부터 최종 적합도를 제공한다. 이는 1.3.2절의 식 (1.8)의 $ST2$ 공식에 의한 것으로 [프로그램 5.1]의 PROC MDS 문의 옵션 〈fit=log〉에 대한 디폴트 스트레스 공식이 크루스칼의 스트레스 2이며, 실제 〈fit=log formular=2〉를 지정한 것과 동일하다. 따라서 이들의 평균스트레스 값

$$\frac{0.84 + 0.78 + \ \ldots \ + 0.92}{10} \simeq 0.85$$

가 맨 아래 〈-ALL-〉에 있음을 알 수 있다. 이 값은 차원수 결정을 위한 크루스칼의 판별기준에 따르면 나쁨(0.20)보다 매우 좋지 않아 2차원의 최대우도 다차원척도법의 결과가 자료를 요약하기에 매우 부족하다고 여겨진다. 이를 결과 ⑦ 상그림도 반영하고 있다. ⑥ 차원계수그림에서 보면 2차원의 경우 주체 4와 10은 나머지 사람들과 다르게 가중치가 0.00으로 다차원척도그림의 2차원에 형성된 레크리에이션 간에는 유사하다고 여기고 있다. 1차원의 경우 이들은 가중치가 모두 1.41로 나머지 사람에 비해 훨씬 크므로 1차원에 형성된 레크리에이션 간에 상대적으로 매우 비유사하다고 판단하고 있다.

5.3 삼원과 최대우도 다차원척도법의 비교

삼원과 최대우도 다차원척도법은 공통적으로 $d_{ij}^{(s)}$는 $s(=1,\ldots,m)$번째 주체에 대해 i번째와 j번째 개체 사이의 가중유클리드거리에 의한 비유사성이라고 하면 각 주체에 의해서 삼원행렬 형태인 비유사성행렬 $D^{(s)} = (d_{ij}^{(s)}),\ i,j = 1,\ldots,n$에 대한 다차원척도법이다. 다만 최대우도 다차원척도법의 식 (5.2)에서 정의한 개인지각공간에서 i번째와 j번째 개체 간의 비유사성

$$\delta_{ij}^{(s)} \ = \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)^t W_s (\boldsymbol{x}_i - \boldsymbol{x}_j)\right]^{1/2}$$

은 가중치행렬로 인하여 일반화 가중유클리드거리라고 볼 수 있다. 만약에 가중치행렬을 대각행렬

$$W_s = diag(w_{11}^{(s)} = w_{s1}, ..., w_{kk}^{(s)} = w_{sk})$$

로 한다면 삼원 다차원척도법의 식 (4.1)의 비유사성

$$\delta_{ij}^{(s)} = \left[\sum_{r=1}^{k} w_{sr} (x_{ir} - x_{jr})^2 \right]^{1/2}$$

가 됨을 쉽게 알 수 있다. 그러나 최대우도 다차원척도법에서는 로그-정규분포 $\log d_{ij}^{(s)} \sim N(\log \delta_{ij}^{(s)},\ \sigma_{ij(s)}^2)$를 가정하고 [알고리즘 4]와 같이 최대우도추정의 과정을 거쳐 $k(\leq p)$ 차원의 형상공간좌표를 구하게 된다. 반면에 삼원 다차원척도법인 경우 분포에 대한 가정이 없이 교대최소제곱법에 의해 다차원척도법을 수행하므로 상대적으로 탐색적 다차원척도법으로 여길 수 있다.

삼원과 최대우도 다차원척도법을 위한 각각의 [프로그램 4.2]와 [프로그램 5.1]의 차이점을 살펴보자. PROC MDS 문에서 삼원의 경우는 옵션 〈level=ordinal condition=row coef=diagonal〉를 사용하고 최대우도는 〈fit=log level=loginterval coef=diagonal〉을 사용한다는 점이다. 삼원을 위한 옵션 〈condition=row〉는 입력자료의 형태에 따른 것으로, 최대우도의 경우 이에 해당하는 옵션 〈condition=matrix shape=triangular〉는 디폴트이므로 생략되어 있다. 그리고 [결과 4.2]와 [결과 5.1]의 차이점은 [결과 5.1]의 ③ 변환함수의 기울기와 멱승을 제외하고는 동일하다. 이는 최대우도 다차원척도법에서는 단조함수에 대한 변환을 고려하기 때문이다. 다음으로 두 다차원척도법의 다차원척도그림을 비교하기 위하여 〈그림 5.1〉을 살펴보자. 그림 (b)는 이미 1.1.5절의 [사례 5]에서 최대우도 다차원척도법의 이해를 위해 제공되었고 이를 재인용하였다. 다만 비교를 위해 제공된 (a) 삼원 다차원척도그림은 [프로그램 5.1]의 PROC MDS 문에서 최대우도 다차원척도법의 옵션 〈fit=log level=loginterval coef=diagonal〉 대신에 삼원 다차원척도법의 옵션 〈level=ordinal coef=diagonal〉을 사용하면 된다.

이들 2차원의 다차원척도그림의 최종 적합도는 (a) 삼원 다차원척도그림은 0.32로 상대적으로 (b) 최대우도 다차원척도그림의 0.85에 비해 훨씬 좋은 적합도를 보여 준다. 그러나 차원수 결정을 위한 크루스칼의 판별기준에 따르면 나쁨(0.20)보다 좋지 않아 모두가 2

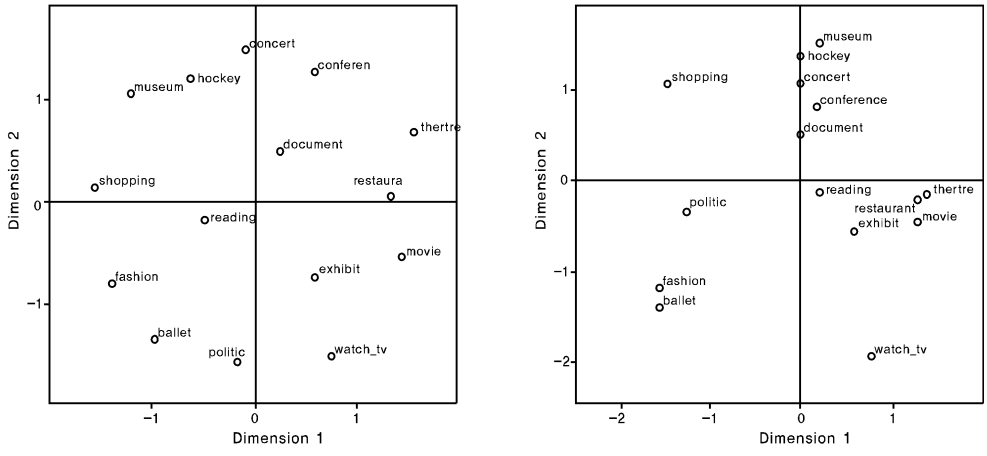
차원의 다차원척도법의 결과가 자료를 요약하기에 매우 부족하다고 여겨진다. 다만 상대적으로 삼원 다차원척도법이 적합도 관점에서 훨씬 안정되어 있다고 볼 수 있다.

[프로그램 5.2] 레크리에이션 자료에 대한 삼원 다차원척도법

```
data recreat;
        input (concert museum theatre movie watch_tv conferen
                reading hockey ballet politic fashion document
                exhibit shopping restaura ) (3.0) @45;
        if _n_ < 16 then subject= 1;
   else if _n_ < 31 then subject= 2;
   else if _n_ < 46 then subject= 3;
   else if _n_ < 61 then subject= 4;
   else if _n_ < 76 then subject= 5;
   else if _n_ < 91 then subject= 6;
   else if _n_ < 106 then subject= 7;
   else if _n_ < 121 then subject= 8;
   else if _n_ < 136 then subject= 9;
   else subject=10;
   cards;
     0
     8   0
    18  21   0
     5  22  16   0
     8  24  22  11   0
    11  13  19  18  22   0
    19  12  15   9  21  14   0
     6   6  15  19  21  13  23   0
    24  18  21  15   9  19  16  23   0
    21  21  23  23  23  13  12  11  23   0
    22  20  24  21  23  10  13  12   7   7   0
    10   7  21  11   9  10   6   8  14  22  22   0
    23  21  18   9   8  10   7  14  22   9  11   8   0
     5  14  19  16  23  22  19  20   8  17  23  21  20   0
    14  23  11   8  21  22   7  10  23  19  21   9   7  22   0
```

```
    0
   22   0
   25   9   0
    3  22  25   0
   14  25  16   7   0
   18  22   7  20  24   0
   22   5  11  14  22   5   0
    9   6  11  22  22   4  19   0
   25  21  18   4  13  18  14  22   0
   22  21  23  16  19  20  21  20  12   0
   19  22  23  21  23   7  21  15   7  18   0
    4  19  23  21  12  22   7   6  19  21  12   0
    9   6  16  12  18  22  22  13  21  20  21   6   0
   21   6  22  22   5  23   4  22  17  19  23  20  17   0
   16  22  17   4   7   5  16  16   6   6  21   5  23  17   0
   ;
   run;

 ods graphics on;
 proc mds data=recreat fit=log level=oridnal coef=diagonal
pfinal nophist;
       subject subject;
```

그림 (b)에서 수평축인 1차원축(Dimension 1)은 왼편의 쇼핑, 정치적 논쟁, 패션쇼, 발레 등과 같이 현장감이 있는 동적인 레크리에이션이며 오른편의 독서, 영화, 박물관, 전시회 등은 주로 정적인 레크리에이션으로 구분하고 있다.



(a) (b)

〈그림 5.1〉 삼원과 최대우도 다차원척도그림: (a) 삼원 다차원척도그림, (b) 최대우도 원척도법(재인용)

반면에 (a)에서는 이미 (b)에서 1차원축에 대하여 동적인 레크리에이션 경향을 보인 쇼핑, 정치적 논쟁, 패션쇼, 발레 외에도 (b)의 1차원축의 중앙인 원점 주위에 위치한 박물관, 하키관람, 독서, 콘서트가 동적인 레크리에이션으로 분류되고 있다. 독서가 의외이기는 하지만 나머지는 정적인 것으로 보기보다는 현장에서 느끼는 레크리에이션으로 충분히 여길 만하다. 사실 (b)에서 기하적으로 1차원축의 중앙 원점 주위에 위치한 레크리에이션들은 어느 한쪽의 경향도 뚜렷하게 나타내고 있지 않아 정적인 것으로 분류하는 것도 애매한 면은 있다. 이는 최대우도 다차원척도법의 낮은 적합도 문제와도 관련성이 있다고 여겨지며 아무튼 큰 맥락에서는 서로 유사한 경향을 보여 준다.

연습문제

5.1 다음은 18명의 학생들로부터 9점 척도[매우 다르다(1점) ~ 매우 유사하다(9점)]로 얻어진 12개 국가의 유사성 자료 중 2명의 학생에 대한 자료이다(Kruskal과 Wish, 1978, p. 64). 밑쪽 삼각행렬자료는 학생 1에 대한 유사성행렬이고 위쪽 삼각행렬자료는 학생 2에 대한 유사성행렬이다. 이들 18명 학생들의 응답의 평균으로 된 자료는 이미 1.1.3절 [사례 3]에서 살펴본 것이다. [부록 2: 자료 9(파일이름: nation.txt)]를 참고.

Brazil	.	6	8	5	3	4	3	3	2	2	5	2
Congo	7	.	6	7	3	5	4	3	2	2	2	2
Cuba	8	5	.	4	2	3	2	3	3	2	3	2
Egypt	2	7	6	.	2	3	5	2	2	2	2	4
France	3	3	2	6	.	3	4	3	2	4	8	5
India	2	7	2	6	2	.	3	5	5	4	5	4
Israel	3	2	3	8	5	7	.	3	3	4	6	7
Japan	2	1	1	4	3	7	2	.	7	6	7	4
China	2	6	8	7	4	5	3	7	.	8	3	4
USSR	2	3	7	7	3	6	3	6	8	.	5	6
USA	5	3	3	4	8	5	7	8	1	2	.	3
Yugoslavia	2	2	8	3	4	5	4	6	8	8	2	.

1) 최대우도 다차원척도법을 실시하고 해석하라.
2) 적절한 차원수를 정하고 적합도에 대하여 해석하라.
3) 삼원 다차원척도법과 비교하라.
4) 차원계수그림을 통하여 비유사성을 각 주체들이 판단하는 정도를 해석하라.

5.2 4장의 [연습문제 4.1]에서 삼원 다차원척도법을 적용한 8개 나라 비유성자료에 대하여 최대우도 다차원척도법을 실시하고 다차원척도그림과 상그림을 서로 비교하라.

5.3 4장의 [연습문제 4.2]의 4가지 식용작물의 비유사성자료에 대해 최대우도 다차원척도법을 실시하고 삼원 다차원척도법의 결과와 비교하라.

5.4 4장의 [연습문제 4.3]의 8가지 관심 영역(건강, 과학, 기술, 상업, 사무, 중개, 사회, 예술)에 대한 4개 그룹(유색 여성, 아시아 여성, 남아프리카 공용 네덜란드어를 사용하는 백인 여성, 영어를 사용하는 백인 여성) 각각의 심리적 시험 점수의 상관계수에 대한 자료를 고려하자.

1) 스트레스-차원수 그림을 통하여 적절한 차원수를 결정하라.
2) 최대우도 다차원척도법을 실시하고 상그림을 통하여 모형의 적합성을 논의하라.
3) 최대우도 다차원척도그림을 제공하고 해석하라.
4) [연습문제 4.3]의 삼원 다차원척도법의 결과와 비교하라.
5) 차원계수그림을 통하여 8개 관심영역에 대한 각 주체들이 판단하는 정도를 해석하라.

6장
베이즈 다차원척도법

다차원척도법에서 베이즈 추정을 적용한 베이즈 다차원척도법(Bayesian MDS)에 대한 모형을 Oh와 Raftery(2001)가 개발하여 소개하였다. 특히, 그들은 계량형 다차원척도법에서 MCMC(Markov chain Monte Carlo) 알고리즘에 바탕을 둔 베이즈 방법을 적용하여 저차원 형상공간을 얻고 차원수를 결정하는 방법을 제공하였는데 이들의 연구를 보다 엄밀히 말하자면 베이즈 계량형 다차원척도법을 개발하였다. 이들의 연구는 기존의 방법보다 자료를 더 잘 적합시키고 측정오차와 어긋난 모형의 가정에 로버스트함을 보였다. 지금까지 살펴본 다차원척도법에서는 대부분이 고차원의 측정 자료로부터 얻어지는 개체 간의 비유사성 d_{ij}와 차원축소를 통해 얻어지는 저차원의 형상좌표자료의 비유사성 δ_{ij}와의 관계에서 측정오차 ϵ_{ij}를 고려하고 있다. 다차원척도법에서 이 측정오차에 대하여 베이즈 관점과 같은 모형 틀을 고려하고자 하는 것이다.

따라서 베이즈 다차원척도법과 관련된 베이즈 모형을 소개하고 사례를 통하여 기존 계량형과 베이즈 계량형 다차원척도법을 서로 비교하고 그 차이점을 이해하려고 한다. 그리고 베이즈 계량형 다차원척도법을 위한 R-프로그램과 그 결과를 제시하여 활용도를 높이고자 한다.

6.1 베이즈 계량형 다차원척도법

베이즈 계량형 다차원척도법은 Oh와 Raftery(2001)가 제안하였다. 일반적으로 통계적 모형에서 사용될 분포의 종류를 가정하기가 가끔 쉽지는 않다. 베이즈 계량형 다차원척도법도 예외는 아니다. Okada와 Shigemasu(2009, 2010), Okada와 Mayekawa(2011), Okada(2012)도 다양한 상황에서 Oh와 Raftery(2001)의 베이즈 추정법을 적용한 다양한 베이즈 다차원척도법을 소개하고 있다. 특히, Okada와 Shigemasu(2009)는 베이즈 다차원척도법의 알고리즘에 대하여 Oh와 Raftery(2001)의 FORTRAN 프로그램을 R-프로그램으로 전환하여 제공하고 있다. 이 절에서는 베이즈 다차원척도법을 위한 적절한 분포를 가정하고 사전과 사후 분포의 관계를 통해 수치해석 관점에서 몬테칼로방법인 MCMC 알고리즘과 이를 위한 임의보행 메트로폴리스-해스팅스(random walk Metropolis-Hastings) 알고리즘에 의한 베이즈 추정법에 대하여 알아보고자 한다.

6.1.1 베이즈 모형

다변량자료행렬 X의 i번째 행벡터 $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ip})^t$는 관측된 i번째 개체를 나타내며 이들 n개의 행벡터 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$은 p차원의 공간에서 n개 개체들의 좌표점이다. 식 (1.3)으로부터 두 개체 $\boldsymbol{x}_i$와 $\boldsymbol{x}_j$ 사이의 유클리드거리

$$d_{ij} \;=\; \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)^t(\boldsymbol{x}_i - \boldsymbol{x}_j)\right]^{1/2} \;=\; \left[\sum_{r=1}^{p} (x_{ir} - x_{jr})^2\right]^{1/2}$$

를 고려하여 이를 p차원의 유클리드공간에서 두 개체 간의 비유사성으로 하자. 따라서 일반적으로 다차원척도법이란 비유사성으로 정의한 거리 d_{ij}와 차원 축소된 $k(\le p)$ 차원의 공간에서 두 개체 사이의 거리 δ_{ij} 사이의 관계가 서로 일치되도록 하는 것이다. 여기서 이들 간의 관계는 일반적으로 1.3.1절의 측정수준에서 4가지 모형으로 임의의 함수 $f(\cdot)$에 대해서

$$d_{ij} = f(\delta_{ij}) + \epsilon_{ij},\;\; i,j = 1,\ldots,n$$

로 정의되며, 여기서 ϵ_{ij}는 측정 및 $k(\le p)$차원 공간근사에 따른 왜곡오차인 오차항이다.

특별히 베이즈 다차원척도법의 모형을 위해서 함수 $f(\cdot)$를 절대척도 모형인 $f(\delta_{ij}) = \delta_{ij}$로 두자. 그리고 오차항에 대하여 다음의 정규분포

$$\epsilon_{ij} \sim N(0, \sigma^2), \ i, j = 1, \ldots, n \tag{6.1}$$

를 가정하자.

베이즈 다차원척도법에 의해서 얻고자 하는 $k(\leq p)$차원 형상좌표행렬 $X_{(k)} = (x_{ir})$, $i = 1, \ldots, n; r = 1, \ldots, k$에서 x_{ir}은 비관측(unobserved)좌표이며 i번째 행벡터 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^t$로부터 형상공간의 두 개체 간의 비유사성은 다음과 같이 정의된다.

$$\delta_{ij} = \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)^t (\boldsymbol{x}_i - \boldsymbol{x}_j) \right]^{1/2} = \left[\sum_{r=1}^{k} (x_{ir} - x_{jr})^2 \right]^{1/2}$$

이 형상공간의 비유사성 δ_{ij}를 미지(unknown)라고 하고, p차원 유클리드공간의 비유사성 d_{ij}가 절단(truncated)정규분포

$$d_{ij} \sim N(\delta_{ij}, \sigma^2) I_{[d_{ij} > 0]}, \ i \neq j, \ i, \ j = 1, 2, \cdots, n \tag{6.2}$$

를 따른다고 가정하자. 식 (6.2)의 가정으로부터 비유사성행렬 $D = (d_{ij})$가 주어질 때 k차원의 형상좌표행렬 $X_{(k)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^t$와 σ^2은 미지의 모수로 이들에 대한 우도함수는 다음과 같다.

$$\begin{aligned} L\left(X_{(k)}, \sigma^2 \mid D\right) &= \prod_{i<j} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(d_{ij} - \delta_{ij})^2}{2\sigma^2} \right\} \left\{ 1 - \Phi\left(-\frac{\delta_{ij}}{\sigma} \right) \right\}^{-1} \\ &\propto (\sigma^2)^{-m/2} \exp\left\{ -\frac{RSS}{2\sigma^2} - \sum_{i<j} \log \Phi\left(\frac{\delta_{ij}}{\sigma} \right) \right\}, \end{aligned} \tag{6.3}$$

여기서 $RSS = \sum\sum_{i<j} (d_{ij} - \delta_{ij})^2$는 식 (1.5)에서 정의한 잔차제곱합이며, $\Phi(\cdot)$는 표준정규분포의 누적함수이고 $m = n(n-1)/2$은 비유사성의 개수이다.

식 (6.3)으로부터 베이즈 추정을 위해 k차원 형상좌표행렬 $X_{(k)}$와 σ^2에 대한 사전(prior)분포가 요구된다. 먼저 $X_{(k)}$의 i번째 행 $\boldsymbol{x}_i$, $i=1,...,n$의 사전분포로 평균이 $\mathbf{0}$이고 대각 공분산행렬 $\Sigma_{xx}=diag\{\lambda_1,\ \cdots\ ,\ \lambda_k\}$를 가지는 다변량정규분포

$$\boldsymbol{x}_i \sim N_k(\mathbf{0}, \Sigma_{xx}),\ i=1,...,n \tag{6.4}$$

를 따른다고 하자.

$X_{(k)}=(\boldsymbol{x}_1,...,\boldsymbol{x}_n)^t$의 완전조건부 사후(posterior)밀도함수는

$$\pi(X_{(k)}|\sigma^2,\lambda_r)\ \propto\ (\sigma^2)^{-m/2}\left(\prod_{r=1}^{k}\lambda_r^{-n/2}\right)\exp\left\{-\frac{Q_1+Q_2}{2}-\sum_{i<j}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right\} \tag{6.5}$$

이다. 여기서 $Q_1=RSS/\sigma^2$과 $Q_2=\sum_{i=1}^{n}\left(\boldsymbol{x}_i^t\Sigma_{xx}^{-1}\boldsymbol{x}_i\right)=\sum_{r=1}^{k}\lambda_r^{-1}\sum_{i=1}^{n}x_{ir}^2$는 $X_{(k)}$의 이차(quadratic)함수이다.

다음으로 오차항의 분산 σ^2에 대한 켤레(conjugate)사전분포로

$$\sigma^2\ \sim IG(a,\ b) \tag{6.6}$$

을 고려하자. 여기서 $IG(a,\ b)$는 모수 a와 b를 가진 역-감마분포로 $\pi(\sigma^2)\propto(\sigma^2)^{-(a+1)}e^{-b/\sigma^2}$, $a,\ b>0$를 따른다. 유사하게 λ_r의 사전분포로

$$\lambda_r\ \sim IG(\alpha,\ \beta_r), r=1,...,k \tag{6.7}$$

로 두면 $\pi(\lambda_r)\propto(\lambda_r)^{-(\alpha+1)}e^{-\beta_r/\lambda_r}$, $\alpha,\ \beta_r>0$이다. 끝으로 $X_{(k)}$, $\{\lambda_r\}$, 그리고 σ^2의 사전분포들은 $\pi\left(X_{(k)},\ \{\lambda_r\},\ \sigma^2\right)=\pi(X_{(k)})\cdot\pi(\sigma^2)\cdot\pi(\{\lambda_r\})$을 만족하는 서로 독립이라고 가정하자.

그러면 식 (6.2)의 가정으로부터 비유사성행렬 $D=(d_{ij})$가 주어질 때 미지의 모수 $(X_{(k)},\ \{\lambda_r\},\ \sigma^2)$의 결합사후밀도함수는 식 (6.3)의 우도함수와 사전분포함수 식 (6.4), 식 (6.6) 그리고 식 (6.7)로부터

$$\pi\left(X_{(k)},\ \{\lambda_r\},\ \sigma^2 \mid D\right) = L\left(X_{(k)},\ \sigma^2 \mid D\right)\cdot \pi\left(X_{(k)}\right)\cdot \pi(\sigma^2)\cdot \pi\left(\{\lambda_r\}\right),$$
$$\propto (\sigma^2)^{-(m/2+a+1)}\left(\prod_{r=1}^{k}\lambda_r^{-(n/2+\alpha+1)}\right)e^{-Q} \tag{6.8}$$

이다. 여기서 Q는 다음과 같다.

$$Q = \frac{Q_1+Q_2}{2} + \sum_{i<j}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right) + \frac{b}{\sigma^2} + \sum_{r=1}^{k}\frac{\beta_r}{\lambda_r}.$$

6.1.2 MCMC 알고리즘

비유사성행렬 $D=(d_{ij})$가 주어질 때 미지의 형상좌표행렬 $X_{(k)}$, 모수 $\{\lambda_r\}$과 σ^2의 결합사후밀도함수인 식 (6.8)의 복잡한 형태 때문에 모수들의 베이즈 추정치를 얻기 위해서는 수치적분(numerical integration)이 요구된다. 이를 위해서 MCMC 추정법을 사용하며 사후밀도함수의 $X_{(k)}$와 미지의 모수 $\{\lambda_r\}$과 σ^2의 초깃값으로 기존의 토거선의 알고리즘에 바탕을 둔 계량형이나 ALSCAL 알고리즘에 바탕을 둔 삼원 다차원척도법의 결과를 활용한다.

먼저 이들 기존의 다차원척도법으로부터 얻어진 형상좌표행렬을 중심화를 시켜 초깃값으로 $X_{(k)}^{(0)}$라고 하면 행들의 합은 $\sum_{i=1}^{n}\boldsymbol{x}_i^{(0)}=0$을 만족한다. 이로부터 표본 공분산행렬 $S_{xx}^{(0)} = X_{(k)}^{(0)t}X_{(k)}^{(0)}/n = \sum_{i=1}^{n}\boldsymbol{x}_i^{(0)t}\boldsymbol{x}_i^{(0)}/n$를 구하게 된다. 이때 $S_{xx}^{(0)}$의 $r(=1,\ldots,k)$번째 대각원소 $s_{rr}^{(0)}$은 분산이며 λ_r의 초깃값 $\lambda_r^{(0)} = s_{rr}^{(0)}$로 사용할 수 있다. 더군다나 중심화된 $X_{(k)}^{(0)}$로부터 잔차제곱합 RSS를 계산하여 이를 초깃값 $RSS^{(0)}$라고 하면 $\sigma^{2(0)} = RSS^{(0)}/m$을 오차항의 분산 σ^2의 초깃값으로 사용할 수 있다.

이번에는 MCMC 알고리즘에서 표본생성에 대하여 상세한 설명을 하기로 하자. 각 반복단계에서 식 (6.8)의 비유사성행렬 $D=(d_{ij})$가 주어질 때 미지의 모수 $(X_{(k)},\ \{\lambda_r\},\ \sigma^2)$의 결합사후밀도함수 $\pi\left(X_{(k)},\ \{\lambda_r\},\ \sigma^2 \mid D\right)$로부터 λ_r의 완전조건부 사후분포는

$$\pi(\lambda_r|X_{(k)},\sigma^2,D) \sim IG\left(\alpha+\frac{n}{2},\ \beta_r+\frac{s_r}{2}\right), r=1,\ldots,k \tag{6.9}$$

이 된다. 여기서 $s_r = \sum_{i=1}^{n} x_{ir}^2 = ns_{rr}$이다.

이번에는 MCMC 알고리즘 각 단계에서 $X_{(k)}$의 i번째 행 $\boldsymbol{x}_i$와 σ^2을 생성하기 위해서 Hastings(1970)의 임의보행 메트로폴리스-해스팅스 알고리즘을 적용하였다. 식 (6.8)로부터 $\boldsymbol{x}_i$의 완전조건부 사후분포는

$$\pi\left(\boldsymbol{x}_i \mid \{\lambda_r\},\ \sigma^2,\ D\right) \propto \exp\left[-\frac{1}{2}(q_1 + q_2) - \sum_{j \neq i, j=1}^{n} \log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right] \tag{6.10}$$

이다. 여기서 $q_1 = \sum_{j \neq i, j=1}^{n} \left(d_{ij} - \delta_{ij}\right)^2 / \sigma^2$과 $q_2 = \sum_{r=1}^{k} x_{ir}^2 / \lambda_r$이다.
다음으로 σ^2의 완전조건부 사후분포는 근사적으로

$$\sigma^2 \sim^{App.} IG\left(\frac{m}{2} + a,\ \frac{RSS}{2} + b\right) \tag{6.11}$$

을 따르게 된다. 여기서 $m = n(n-1)/2$은 비유사성의 개수이다. 일반적으로 개체수 n의 이차함수인 m의 값이 크게 되므로 식 (6.11)의 역-감마분포는 정규분포로 잘 근사된다. 따라서 식 (6.11)의 역-감마분포의 분산에 비례하는 분산을 가진 정규분포로 임의보행 메트로폴리스-해스팅스 알고리즘을 제안하게 된다.

형상좌표행렬 $X_{(k)}$, 미지의 모수 $\{\lambda_r\}$과 σ^2의 충분한 시간에 걸친 반복 생성은 미지의 모수에 의한 사후분포를 통해 표본을 얻고 이로부터 모수들의 베이즈 추정을 하게 된다. 일반적으로 다차원척도법과 관련된 모형에서는 형상공간의 미지의 비유사성 δ_{ij}가 유클리드 거리로 정의되어 있으므로 형상좌표행렬 $X_{(k)}$의 사후표본은 강력한 사전정보가 없는 한 다른 다차원척도법과 같이 원점에 대하여 이동, 회전 그리고 반사에도 불변하게 된다. 따라서 MCMC 알고리즘의 수렴을 검증하기 위해서는 $X_{(k)}$보다는 비유사성행렬 D의 수렴을 점검해야 한다.

더군다나 Okada와 Mayekawa (2011)는 손실함수에 의한 사후-과정(post-processing) 알고리즘을 제안하여 객관적 수렴과정을 점검하였다. $X^{(i)}$를 $X_{(k)}$의 i번째 MCMC 표본이라고 하고, $\overline{X}$는 모든 $X^{(i)}$들의 평균이면서 프로크러스티즈(Procrustes)회전의 목표행렬이

라고 하고, $H^{(i)}$를 $X^{(i)}$에 대응하여 추정될 회전행렬이라 하자. 사후-과정은 손실함수

$$\rho = (\overline{X} - X^{(i)}H^{(i)})^t (\overline{X} - X^{(i)}H^{(i)}) \tag{6.12}$$

를 최소화하는 것으로 다음과 같이 [알고리즘 5]에 요약할 수 있다.

[알고리즘 5] 사후-과정 알고리즘
1단계: MCMC 표본 $X^{(i)}$들을 마련하고 평균 $\overline{X}$를 구한다. **2단계:** 프로크러스티즈회전에 의해서 목표행렬 $\overline{X}$로 각 표본 $X^{(i)}$를 회전시킨다. **3단계**: 새롭게 회전하여 얻어진 $X^{(i)}$들로부터 $\overline{X}$를 다시 계산한다. **4단계:** 수렴할 때까지 [2단계]와 [3단계]를 반복한다.

6.1.3 차원의 수

일반적으로 다차원척도법에서는 궁극적으로 개체들의 관계를 보여 주는 형상공간의 차원의 수 $k(\leq p)$는 미지이고 이를 결정하는 것이 중요하다. 이미 1.3.3절에서 스트레스를 최소화하는 관점에서 스트레스-차원수 그림이나 크루스칼 판별기준을 제시한 바 있다. Oh와 Raftery(2001)는 베이즈 다차원척도법에서도 차원의 수를 최종적으로 얻어지는 개체 형상공간에 바탕을 두는 간단한 베이즈 차원 선택법을 제안하고 있다.

차원 k를 미지의 변수라고 하면 사후분포는 식 (6.8)로부터 다음과 같다.

$$\begin{aligned}\pi\left(X_{(k)},\ \{\lambda_r\},\ \sigma^2, k|\ D\right) &\propto L\left(X_{(k)},\ \sigma^2, k|\ D\right)\cdot\pi\left(X_{(k)}|\{\lambda_r\}, k\right)\cdot\pi(\sigma^2)\cdot\pi\left(\{\lambda_r\}|k\right)\\ &= A(k)\ \bullet\ h(\sigma^2, X_{(k)})\ \bullet\ g(\{\lambda_r\}, X_{(k)}, k).\end{aligned} \tag{6.13}$$

여기서

$$A(k) = (2\pi)^{-(m+nk)/2}\Gamma(a)^{-1}b^a\Gamma(\alpha)^{-k}\prod_{r=1}^{k}\beta_r^{\alpha},$$

$$h(\sigma^2, X_{(k)}) = (\sigma^2)^{-(m/2+a+1)}\exp\left[-\frac{(RSS/2+b)}{\sigma^2} - \sum_{i<j}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right],$$

$$g(\{\lambda_r\}, X_{(k)}, k) = \prod_{r=1}^{k} \lambda_r^{-(n/2+\alpha+1)} \exp\left[-\frac{(s_r/2+\beta_r)}{\lambda_r}\right]$$

이다.

식 (6.13)으로부터 주변사후분포함수는

$$\begin{aligned}\pi\left(X_{(k)},\ k|\ D\right) &= c \bullet A(k) \bullet \int h(\sigma^2, X_{(k)})d\sigma^2 \bullet \int g(\{\lambda_r\}, X_{(k)}, k)d\{\lambda_r\} \\ &= c \bullet A^*(k) \bullet \left(\frac{RSS}{m}\right)^{-m/2+1} \prod_{r=1}^{k} \left(\frac{s_r}{2} + \beta_r\right)^{-(n/2+\alpha)} \qquad (6.14)\end{aligned}$$

이다. 여기서 $A^*(k) = A(k) \bullet (2\pi)^{1/2}\Gamma^k(n/2+\alpha)(m/2)^{-1/2}e^{-m/2}$이고 c는 $X_{(k)}$와 k와는 독립인 상수이다.

식 (6.14)의 $\prod_{r=1}^{k}(s_r/2+\beta_r)^{-(n/2+\alpha)}$에서 $s_r/n = s_{rr}$로 $X_{(k)}$의 r번째 좌표의 분산이므로 $\pi\left(X_{(k)},\ k|\ D\right)$는 $X_{(k)}$의 척도(scale)에 종속되어 있어 이 분산이 작을수록 더 커지게 된다. 즉, $X_{(k)}$의 척도는 차원의 수 k에 따라 변할 수 있다. 만약에 다음과 같이 1차원과 2차원의 공간좌표가 있다고 하면

$$X_{(1)} = (-1\ \ 1),\ X_{(2)} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

이들 공간좌표는 유클리드거리가 4로 동일하나 2차원의 좌표 $X_{(2)}$의 분산이 더 작게 된다. 따라서 이러한 차원에 따른 척도 문제를 극복하기 위해서 $X_{(k)}$대신에 처음 k개 좌표는 $X_{(k)}$와 같고 나머지 좌표는 모두 0인 것으로 하여 $k+1$ 차원의 형상좌표 $X^*_{(k+1)} = (X_{(k)} : \mathbf{0})$를 만들면 된다. 따라서 $X^*_{(k+1)}$와 $X_{(k+1)}$를 비교하여 $X^*_{(k+1)}$가 $X_{(k+1)}$보다 큰 주변사후분포함수값을 가지면 차원의 수로 k를 선택하면 된다.

이를 위하여 식 (6.14)로부터 $X^*_{(k+1)}$와 $X_{(k+1)}$의 주변사후분포함수의 비(ratio)

$$R_{(k)} \equiv \frac{\pi\left(X_{(k+1)},\ k+1|\ D\right)}{\pi\left(X^*_{(k+1)},\ k+1|\ D\right)} \qquad (6.15)$$

를 고려하고 $X_{(k)}$의 r번째 좌표의 분산으로부터 계산된 $s_r = s_r^{(k)}$와 이로부터 $t_r^{(k+1)} = s_r^{(k+1)}/s_r^{(k)}$라고 하고 $\alpha = 1/2$과 $\beta_r = s_r^{(k+1)}/2n$이라 두면 식 (6.15)의 비$R_{(k)}$의 $-2\log R_{(k)}$는

$$
\begin{aligned}
LR_{(k)} &\equiv -2\log R_{(k)} \\
&= (m-2)\log\left(\frac{RSS_{(k+1)}}{RSS_{(k)}}\right) \qquad (6.16) \\
&+ (n+1)\left\{\sum_{r=1}^{k}\log\left[\frac{t_r^{(k+1)}(n+1)}{n+t_r^{(k+1)}}\right]+\log(n+1)\right\}. \qquad (6.17)
\end{aligned}
$$

이 됨을 쉽게 보여 줄 수 있다.

식 (6.16)을 살펴보면 $LR_{(k)}$은 로그-우도비와 유사하고 이는 $k+1$차원의 잔차제곱합 $RSS_{(k+1)}$가 상대적으로 k차원의 $RSS_{(k)}$보다 작을수록 음의 값을 가진다. 식 (6.17)은 차원이 1씩 높아질 때마다 증가되어 벌점(penalty)과 같은 역할을 한다. 만약 $X_{(k)}$와 $X_{(k+1)}$ 사이에 유의한 변화가 없다면 $t_r^{(k+1)} \simeq 1$이 되어 벌점은 근사적으로 $(n+1)\log(n+1)$이 된다.

따라서 $LR_{(k)}$의 값이 양이면 차원의 수는 k이며 음이면 $k+1$을 선호하게 되므로 이 값이 양으로 변하는 곳에서 차원의 수를 결정하면 된다. 대안적으로 다음의 $MDSIC$를 제안하고 있다.

$$
\begin{aligned}
MDSIC_{(1)} &= (m-2)\log RSS_{(1)}, \\
MDSIC_{(k)} &= MDSIC_{(1)} + \sum_{r=1}^{k-1} LR_{(r)}. \qquad (6.18)
\end{aligned}
$$

식 (6.18)의 $MDSIC_{(k)}$를 최소화하는 차원의 수를 최적 차원으로 정하면 된다.

6.2 베이즈 계량형 다차원척도법의 예

6.1절에서 정리한 베이즈 계량형 다차원척도법을 적용하기 위한 사례로 Okada와 Shigemasu(2009)에서 소개한 유럽 10대 도시 간 거리 자료를 이용하려 한다. 기존의 계량형 다차원척도법과의 상세한 비교는 6.3절로 미루고 여기서는 베이즈 계량형 다차원척도법에 대하여 설명하고자 한다. StatLib 사이트(lib.stat.cmu.edu/general/BMDS.f90)에서 Oh와 Raftery(2001)의 베이즈 계량형 다차원척도법을 위한 FORTRAN 프로그램 대신에 최근 Okada와 Shigemasu(2009)가 R과 WinBUG를 사용한 BMDS 프로그램을 참고하려고 한다.

6.2.1 유럽 10대 도시 간의 거리 자료

유럽의 대표적인 10대 도시인 London, Stockholm, Lisboa, Madrid, Paris, Amsterdam, Berlin, Prague, Rome, Dublin 간의 거리에 대한 자료가 〈표 6.1〉에 주어져 있다(Borg와 Groenen, 2005, p. 20).

〈표 6.1〉 유럽 10대 도시 간 거리

London	0									
Stockholm	569	0								
Lisboa	667	1212	0							
Madrid	530	1043	201	0						
Paris	141	617	596	431	0					
Amsterdam	140	446	768	608	177	0				
Berlin	357	325	923	740	340	218	0			
Prague	396	423	882	690	337	272	114	0		
Rome	569	787	714	516	436	519	472	364	0	
Dublin	190	648	714	622	320	302	514	573	755	0

참고로 이 자료는 [부록 2: 자료 12(파일이름: eurodist.txt)]에 분석을 위해 정리되어 있으니 활용하기를 바란다.

6.2.2 R-프로그램과 결과

[프로그램 6.1]의 # Bayes Metric MDS 이하는 베이즈 계량형 다차원척도법을 6.2.1절에서 설명한 자료에 적용하기 위한 R-프로그램으로 bmds 함수를 활용하고 있다. 특히, 이

를 사용하기 위해선 먼저 WinBUGS 프로그램과 BMDS 코드를 BUGS 전용 사이트로부터 내려 받고 WinBUGS 개발 전용 사이트(winbugs-development.org.uk)로부터 WBDev를 다운받아 WinBUGS를 업데이트해야 한다. 다음으로 R2WinBUGS, gtools, plotrix를 R-프로그램의 메뉴 패키지에서 선택하고 bmds함수를 사용하기 전에 library(R2WinBUGS), library(gtools), library(plotrix) 문에 의해서 추가적으로 지정하여 RGui 환경을 조성해야 한다. 그리고 BMDS 코드를 사전에 실행해야 한다(Okada와 Shigemasu, 2009).

[프로그램 6.1] 〈표 6.1〉 유럽 10대 도시 간 거리에 대한 베이즈 계량형 다차원척도법을 위한 R-프로그램

```
# Bayes Metric MDS
delta<-read.table("c:/ProgramFiles/WinBUGS14/eurodist.txt", header=T)
delta
bmds.res<-bmds(delta, maxp=5, n.iter=25000, n.burnin=5000, MCMCsample=T)
bmds.res$MDSIC
bmds.res$stress
plot.bmds(bmds.res, p=2)
abline(v=0, h=0)
```

[프로그램 6.1]에서 먼저 입력자료에 대한 다음의 명령문

delta<-read.table("c:/ProgramFiles/WinBUGS14/eurodist.txt", header=T)

에서 read.table 함수의 〈c:/ProgramFiles/WinBUGS14/eurodist.txt〉는 드라이브 c:에 위치한 폴더 〈ProgramFiles/WinBUGS14〉의 자료 〈eurodist.txt〉를 불러들여 표 형식의 〈delta〉로 저장하라는 명령문이다. 〈header=T〉는 첫 행에 개체를 나타내는 유럽의 대표적인 10대 도시인 London, Stockholm, Lisboa, Madrid, Paris, Amsterdam, Berlin, Prague, Rome, Dublin이 있음을 알리는 인수이다.

베이즈 계량형 다차원척도법을 위한 bmds 함수 bmds(delta, maxp=5, n.iter=25000, n.burnin=5000, MCMCsample=T)를 이용하였고 인수로 maxp=5는 1차원부터 5차원까지 형상공간의 적합을 요구하며, n.iter=25000, n.burnin=5000, MCMCsample=T는 MCMC 알고리즘과 관련된 인수로 수치적 반복수 25,000과 [알고리즘 5]에서 MCMC 표본 5000개를 사용하라는 것이다. 이외에도 인수로 a=5는 식 (6.6)의 오차항의 분산 σ^2에 대한 켤레 사전분포 $\sigma^2 \sim IG(a,\ b)$와 관련되어 있고, alpha=0.5와 beta0=0.5는 식 (6.7)에서 λ_r의

사전분포로 $\lambda_r \sim IG(\alpha,\ \beta_r), r=1,\ldots,k$로부터 식 (6.9)에서 $\alpha=1/2$과 $\beta_r = s_r^{(k+1)}/2n$와 관련된 디폴트 인수이다. [프로그램 6.1]에서 delta, bmds.res$MDSIC, bmds.res$stress, plot.bmds(bmds.res, p=2)는 출력 명령문으로 [결과 6.1]에 순서적으로 정리되어 있다.

먼저 ①은 유럽 10개국의 거리에 의한 비유사성행렬자료로 실제 다차원척도법의 분석에 활용된다. ②에는 6.1.3절의 식 (6.18)에서 최적 차원수를 정하는 $MDSIC_{(k)}$ 통계량값과 이에 관련된 스트레스값이 ③에 제공되고 있다. ④에는 베이즈 계량형 다차원척도그림과 ⑤는 비계량형 다차원척도법에서 제공하는 상그림과 같은 최적 적합의 정도를 나타내는 그림으로 이를 최적적합그림이라고 하자.

[결과 6.1] [프로그램 6.1]의 수행결과: ① 입력자료, ② $MDSIC_{(k)}$ ③ 최종 적합도, ④ 베이즈 계량형 다차원척도그림, ⑤ 최적적합그림

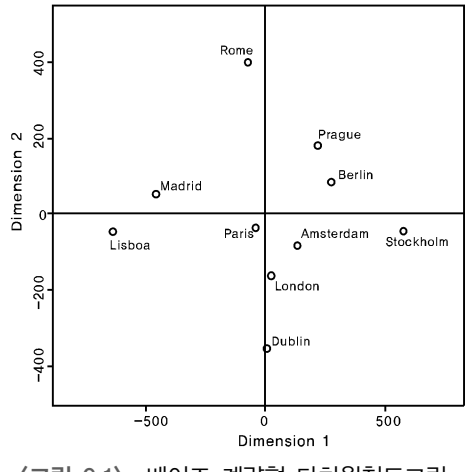
①

	London	Stockholm	Lisboa	Madrid	Paris	Amsterdam	Berlin	Prague	Rome	Dublin
1	0	569	667	530	141	140	357	396	569	190
2	569	0	1212	1043	617	446	325	423	787	648
3	667	1212	0	201	596	768	923	882	714	714
4	530	1043	201	0	431	608	740	690	516	622
5	141	617	596	431	0	177	340	337	436	320
6	140	446	768	608	177	0	218	272	519	302
7	357	325	923	740	340	218	0	114	472	514
8	396	423	882	690	337	272	114	0	364	573
9	569	787	714	516	436	519	472	364	0	755
10	190	648	714	622	320	302	514	573	755	0

② 61.98003 -486.90655 -450.80465 -415.05100 -383.87521

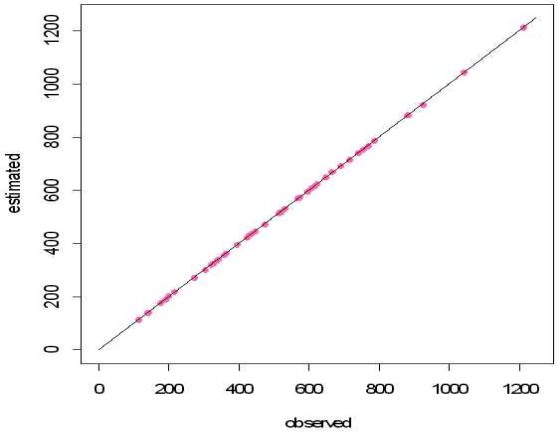
③ 0.3178835892 0.0003956841 0.0004431218 0.0004900834 0.0005143827

④



〈그림 6.1〉 베이즈 계량형 다차원척도그림

⑤



최적적합그림

6.2.3 결과의 해석

[결과 6.1]을 살펴보면 결과 ①은 입력자료로 사용된 유럽 10개국 거리를 비유사성으로 하는 대칭행렬을 나타내고 있다. 베이즈 계량형 다차원 척도법에서 최적의 차원수는 $k=2$인 2차원의 적합이 적절하며 이는 결과 ②에서 식 (6.18)의 $MDSIC_{(k)}$ 통계량값 중에서 최솟값 −486.90은 위치상 두 번째에 해당하기 때문이다. 다음으로 2차원 적합의 정도는 결과 ③과 결과 ⑤로부터 확인 가능하다.

이 경우 결과 ③의 두 번째 스트레스값 0.000396이 충분히 작으며 더군다나 〈표 1.9〉 크루스칼 판별기준에 따라도 〈0.0 완벽(perfect)〉에 가깝다. 결과 ⑤의 최적적합그림은 수직축의 적합된 2차원 공간의 개체 간의 거리 δ_{ij}와 수평축의 비유사성 d_{ij}에 대한 각 개체의 45가지 짝의 산점도이다. 이 그림에서 가장 이상적으로는 좌표점들이 원점을 지나는 직선상에 놓이면 베이즈 계량형 다차원척도법을 위한 베이즈 모형이 적합함을 나타내게 된다. 실제 적합이 매우 잘 된 경우이다. 직선상을 벗어나는 좌표점(이상치)들이 많을수록 스트레스는 커지게 된다. 여기에 대해서는 3장에서 자세히 언급한 바 있다.

결과 ④의 〈그림 6.1〉 베이즈 계량형 다차원척도그림은 유럽 10개국의 지도상의 위치를 상대적으로 잘 나타내고 있다. 다만, 실제 지도를 고려해서 비교한다면 1차원축은 왼편과 오른편이 각각 남북을 나타내며 2차원축의 위쪽과 아래쪽은 각각 동서를 나타내고 있다. 특히, 프랑스 파리는 두 축에 대하여 지리적으로 중간 지점에 위치하고 있다.

6.3 계량형과 베이즈 계량형 다차원척도법의 비교

일반적으로 다차원척도법은 비유사성으로 정의한 거리 d_{ij}와 차원 축소된 $k(\leq p)$차원의 공간에서 두 개체 사이의 거리 δ_{ij} 사이의 관계가 서로 일치되도록 하는 것이다. 여기서 이들 간의 관계는 계량형과 베이즈 다차원척도법의 모형은 다음과 같이 정의하였다.

$$d_{ij} = \delta_{ij} + \epsilon_{ij},\ i,j=1,...,n$$

여기서 ϵ_{ij}는 측정 및 $k(\leq p)$차원 공간근사에 따른 왜곡오차인 오차항이다. 특히, 베이즈 다차원척도법의 모형은 다만 오차항에 대하여 $\epsilon_{ij} \sim N(0,\sigma^2),\ i,j=1,...,n$ 과 같이 정규분포를 가정한다는 점이 계량형 다차원척도법과 다르다. 더불어 p차원 유클리드공간의 비유사성 d_{ij}에 대한 절단(truncated)정규분포 $d_{ij} \sim N(\delta_{ij},\ \sigma^2)I_{[d_{ij}>0]}, i \neq j,\ i,\ j=1,\ 2,\ \cdots,\ n$를 따른다고 가정한다. 더군다나 비유사성행렬 $D=(d_{ij})$가 주어질 때 미지의 $k(\leq p)$차원 형상좌표행렬 $X_{(k)}$와 모수인 $\{\lambda_r\}$과 σ^2의 결합사후밀도함수인 식 (6.8)의 복잡한 형태 때문에 모수들의 베이즈 추정치를 얻기 위한 MCMC 추정법을 사용한다. 이 경우에 수치적 계산을 위해 사후밀도함수의 $X_{(k)}$와 미지의 모수 $\{\lambda_r\}$과 σ^2의 초깃값으로 기존의 토거선의 알고리즘에 바탕을 둔 계량형이나 ALSCAL 알고리즘에 바탕을 둔 삼원 다차원척도법의 결과를 활용한다.

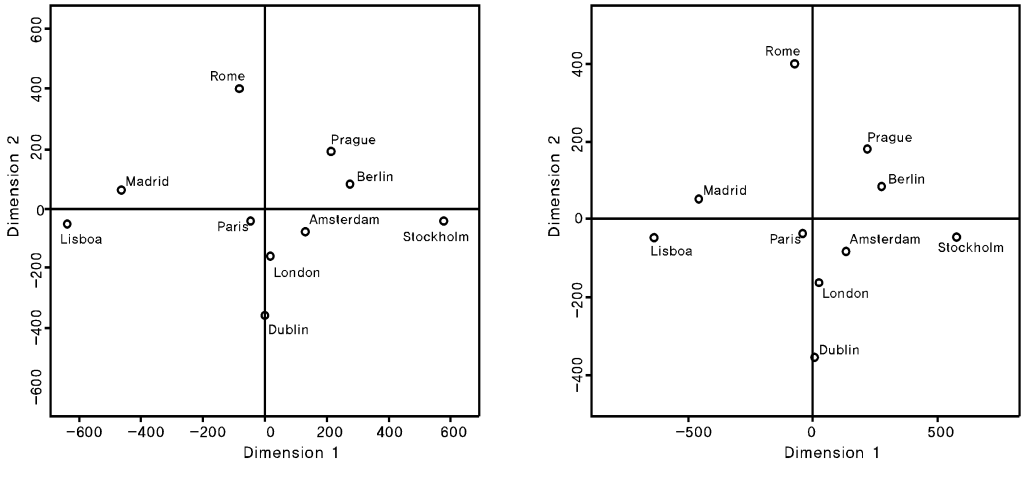
따라서 계량형 다차원척도법을 통해 탐색적으로 개체 간의 비유사성을 살펴본 다음에 이들의 값을 초깃값으로 활용하는 베이즈 다차원척도법을 실시하는 것이 바람직하다.

[프로그램 6.2] 〈표 6.1〉 유럽 10대 도시 간 거리에 대한 계량형 다차원척도법을 위한 R-프로그램

```
# Metric MDS
delta<-as.matrix(read.table( "c:/Program Files/WinBUGS14/eurodist.txt", header=T))
delta
con<- cmdscale(delta, k=2, eig=T)
con
x<-con$points[,1]
y<-con$points[,2]
lim<-c(-max(abs(con$points)), max(abs(con$points)))
plot(x,y, xlab="dimension 1", ylab="dimension 2", xlim=lim, ylim=lim)
text(x,y-20, colnames(delta), cex=0.8)
abline(v=0, h=0)
```

이미 베이즈 계량형 다차원척도법은 [프로그램 6.1]의 수행 결과에서 살펴보았고 계량형 다차원척도법과의 비교를 위하여 다차원척도그림만 재인용하자. [프로그램 6.2]는 베이즈 계량형 다차원척도법을 적용했던 〈표 6.1〉의 유럽 10대 도시 간 거리에 대한 계량형 다차원척도법을 위한 R-프로그램이다. 이는 cmdscale 함수를 사용함을 3장을 통해 잘 알고 있다. 따라서 여기서는 프로그램에 대한 설명은 생략하며 프로그램 실행 결과를 따로 정리 요약하지 않으려 한다. 다만 스트레스값과 다차원척도그림만 비교를 위해 인용하기로 하자.

두 다차원척도법의 비교를 위한 다차원척도그림은 〈그림 6.2〉에 제공되어 있다. 먼저 그들의 최종적합도는 각각 0.999와 0.000396로 2차원 적합이 모두가 완벽에 가까움을 보여 주고 있다. (a) 계량형 다차원척도그림에 대한 경향은 이미 6.2.3절에서 살펴본 (b) 베이즈 계량형 다차원척도그림과 대동소이하다.



〈그림 6.2〉 (a) 계량형 다차원척도그림 (b) 베이즈 계량형 다차원척도그림

연습문제

6.1 다음은 1명의 학생으로부터 9점 척도[매우 다르다(1점) ~ 매우 유사하다(9점)]로 얻어진 12개 국가의 유사성 자료로 그들의 평균에 대한 자료이다(Kruskal과 Wish, 1978, p. 64). 이들 18명 학생들의 응답의 평균으로 된 자료와 2명으로 학생들의 응답 자료는 이미 1.1.3절 [사례 3]에서와 [연습문제 5.1]에서 살펴보았다.

Brazil	.											
Congo	7	.										
Cuba	8	5	.									
Egypt	2	7	6	.								
France	3	3	2	6	.							
India	2	7	2	6	2	.						
Israel	3	2	3	8	5	7	.					
Japan	2	1	1	4	3	7	2	.				
China	2	6	8	7	4	5	3	7	.			
USSR	2	3	7	7	3	6	3	6	8	.		
USA	5	3	3	4	8	5	7	8	1	2	.	
Yugoslavia	2	2	8	3	4	5	4	6	8	8	2	.

1) 베이즈 계량형 다차원척도법을 실시하고 해석하라.
2) 적절한 차원수를 정하고 적합도에 대하여 해석하라.
3) 18명의 평균과 2명의 유사성자료 각각의 계량형과 베이즈 계량형 다차원척도법의 결과와 비교하라.

6.2 Mardia 외 2인(1979, p. 238)에서는 무덤에서 도자기 형태에 대한 고고학에 관련된 자료를 소개하고 있다. G1-G6의 6기 무덤과 그곳에서 소장된 5종류의 도자기에 대하여 i번째 무덤에 j번째 도자기가 있으면 1, 아니면 0으로 하는 이진수자료 형태이다.

무덤	도자기 종류				
	1	2	3	4	5
G1	0	0	1	1	0
G2	1	1	0	0	1
G3	0	1	1	1	1
G4	0	0	1	1	0
G5	1	0	0	0	1
G6	1	0	1	1	1

1) 단순매칭계수를 이용하여 무덤 간의 유사성행렬 C_S를 구하라.

2) 무덤 G1-G6에 대한 $5C_S$는 다음과 같음을 확인하라.

$$5C_S = \begin{pmatrix} 5 & & & & & \\ 0 & 5 & & & & \\ 3 & 2 & 5 & & & \\ 5 & 0 & 3 & 5 & & \\ 1 & 4 & 1 & 1 & 5 & \\ 3 & 2 & 3 & 3 & 3 & 5 \end{pmatrix}$$

3) 유사성 c_{ij}과 비유사성 d_{ij}간의 표준변환 $d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}}$을 이용하여 비유사성행렬을 구하고 토거선 알고리즘에 의한 다차원척도법을 실시하라.

4) 베이즈 계량형 다차원척도법을 실시하고 토거선 알고리즘의 다차원척도법의 결과와 비교하라.

5) 어떤 무덤이 비슷한 종류의 도자기를 소장하고 있는지를 설명하라.

6) 적절한 차원수를 계량형과 베이즈 계량형 다차원척도법에 따라 결정하고 서로 비교하라.

6.3 4장의 [연습문제 4.3]의 8가지 관심 영역(건강, 과학, 기술, 상업, 사무리, 중개, 사회, 예술)에 대한 4개 그룹(유색 여성, 아시아 여성, 남아프리카 공용 네덜란드어를 사용하는 백인 여성, 영어를 사용하는 백인 여성) 각각의 심리적 시험 점수의 상관계수에 대한 자료를 고려하자.

1) 네 그룹별 베이즈 계량형 다차원척도법을 실시하라.

2) 최대우도 다차원척도법의 다차원척도그림과 네 그룹별 베이즈 계량형 다차원척도그림을 비교하고 그 차이점을 논의하라.

6.4 영국의 가장 규모가 크고 오래된 로이드은행(Lloyds Bank)에서 1890 ~ 1970년 사이에 근무한 은행 종사원의 경력에 대한 연구가 Stovel 외 2인 (1996)에 의해서 이루어졌다. 이들은 1900년대 초기에 전통적으로 지위에 근거한 경력제도가 동적인 성취감에 의한 경력제도에 의해서 어떻게 대체되었는지를 살펴보았다. 여기서 제시된 사례에는 로이드은행의 두 개의 코호트인 1905 ~ 1909년과 1925 ~ 1929년 사이에 각각 80명의 은행 종사원의 고용 기록에 대한 자료를 활용하였다. 고용 기록은 각 종사원을 나타내

는 일련번호(id), 처음 근무를 시작한 연도, 그리고 맡은 모든 직무를 1년 단위로 근무한 지점의 규모와 위치(1=작은 지방, 2=큰 지방, 3=작은 도시, 4=큰 도시, 5=전문적 본사나 해외지점, 6=본사)와 맡은 일(1=행원, 2=선임 행원, 3=정규직 경영인, 4=전문 경영인)에 따라 두 자리 숫자를 연속적으로 기록한 것으로 71개 변수가 사용되었다. 이들 지점의 규모와 위치의 조합은 $6 \times 4 = 24$가지로 그 형식은 개인별로 〈해@지점·지위〉로 문자화되었다. 그리고 101-107 숫자를 이용하여 회사로부터 퇴직했음을 표시하였고 은행원의 해마다 경력이 다르므로 필요치 않은 것은 숫자 999로 표시하였다. 물론 이들은 매칭 알고리즘에는 사용되지 않았다. 예를 들면 40년 경력의 일련번호 4238 은행원이 1907년에 업무를 시작하여 25년간 큰 지방지점에서 행원으로 일을 하였고, 15년간 작은 지방지점의 정규직 경영인으로 승진하여 근무하다가 퇴직하였다면 그의 자료는 다음과 같다.

4238 1907 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 13 13 13 13 13 13 13 13 13 13 13 13 13 13 107 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999

실제 이 자료는 개인별 문자화한 형식 〈해@지점·지위〉에 의해서 {25@21, 15@13}와 같이 요약될 수 있다.

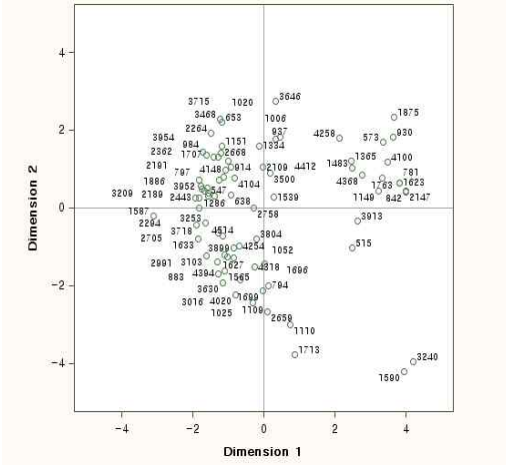
Izenman(2008, 13장)은 이 자료로 서열정렬알고리즘을 통하여 다음과 같은 크기가 80×80 비유클리드거리로 된 비유사사성행렬 자료를 만들었다. 이는 1905 1909년에 근무한 로이드 은행원 80명에 대한 비유사성행렬자료의 일부이다. 515 ~ 4514는 80명 은행원의 일련번호이다. 이와 함께 1925 ~ 1929년의 은행원 80명의 비유사성행렬자료에 대해서는 [부록 2: 자료 3(파일이름: Lloyds05.sd2, Lloyds05.xlsx)]과 [부록 2: 자료 4(파일이름: Lloyds25.sd2, Lloyds25.xlsx]에 SAS와 엑셀자료세트로 각각 제공되어 있으니 참고하기를 바란다.

515	0							
547	4.214285714	0						
573	3.333333333	4.904761905	0					
638	4.108108108	0.892857143	4.837837838	0				
653	4.634146341	2.178571429	4.87804878	2.682926829	0			
781	3.357142857	4.797619048	1.791666667	4.567567568	5.182926829	0		
794	3.511627907	3.406976744	4.953488372	3.697674419	3.279069767	4.965116279	0	
797	4.243902439	0.744047619	4.597560976	1.097560976	1.682926829	4.756097561	·····	
842	3.404761905	4.952380952	1.875	4.878378378	5.158536585	1.045454545	·····	
883	3.795454545	1.886363636	5.102272727	2	3.795454545	4.875	·····	
914	4.31875	1.291666667	4.5875	1.5	1.487804878	4.9875	·····	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
4514	3.464285714	1.238095238	4.836956522	1.488095238	3.059523809	4.797619047	·····	0

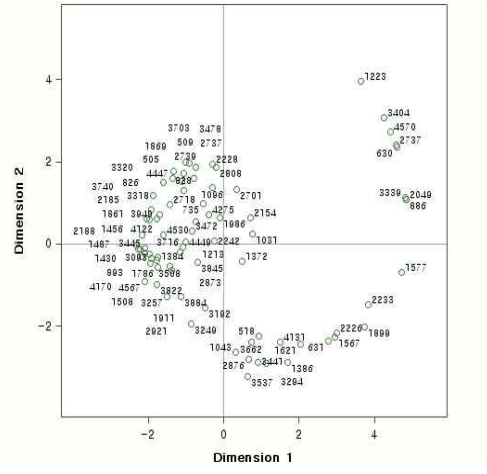
1) 계량형 다차원척도법을 위한 다음의 프로그램을 실행하라.

```
LIBNAME MDS  "C:₩" ;

/* (a) Lloyds Bank Data for 1905-1909 */
data lloyds05;
set mds.lloyds05;
ods graphics on;
proc mds data=lloyds05  level=absolute pfinal plots=coefficients(one);
 id col1;

/* (b) Lloyds Bank Data for 1925-1929 */
data lloyds25;
set mds.lloyds25;
proc mds data=lloyds25  level=absolute pfinal plots=coefficients(one);
 id col1;
```

2) 다음의 다차원척도그림을 얻고 각각의 적합도에 대하여 기술하라.



(a)



(b)

계량형 다차원척도그림: (a) 1905-1909년 (b) 1925-1929년

3) 다차원척도그림 (a)와 (b)를 서로 비교하고, 특히 (a)에서 이상치로 보이는 세 은행원 (1587, 1590, 3240)의 형상좌표를 제시하라.

4) 베이즈 계량형 다차원척도법의 결과와 비교하라.

7장
비선형 다차원척도법

Sammon(1969)의 비선형사상 알고리즘을 적용한 새먼사상(Sammon mapping)은 일반적으로 패턴인식(pattern recognition)에서 잘 알려진 도구이다. 이는 고차원 공간의 개체를 나타내는 벡터를 저차원 공간의 벡터로 사상시키려는 것이다. 물론 개체 간의 패턴이 가능한 유지되도록 하는 것은 당연하다. 이는 다차원척도법의 기하적 관점과 동일하다. 따라서 새먼사상을 다차원척도법에 적용한 것을 비선형 다차원척도법(nonlinear MDS)라 하자. 비선형 다차원척도법은 근본적으로 계량형 다차원척도법의 스트레스 최소화 알고리즘의 비선형 버전으로 생각할 수 있다. 그러므로 엄밀히 말해서 비선형 계량형 다차원척도법이라고 볼 수 있다. 반면에 계량형 다차원척도법을 위한 크루스칼의 스트레스를 최소화하는 알고리즘과 차원축소를 위한 주성분분석의 대수적 성질을 따른 토거선의 알고리즘은 상대적으로 선형 버전이라고 볼 수 있다.
따라서 이 절에서는 다차원척도법의 관점에서 새먼사상의 비선형사상 알고리즘을 간단하게 요약하고, 이를 위해 R-프로그램에서 sammon 함수를 활용하여 사례에 적용하고 기존 계량형 다차원척도법과 비교하려고 한다. 이 함수도 수치적 반복연산을 필요로 하는 비계량형 다차원척도법을 위한 isoMDS 함수처럼 초깃값으로 계량형 다차원척도법을 위한 cmdscale 함수를 사용한다.

7.1 비선형 계량형 다차원척도법

새먼사상은 Sammon(1969)의 비선형사상 알고리즘에 바탕을 둔 것으로, Dzwinel(1994)은 고차원의 자료를 분석하는 데 이것의 유용한 활용성을 요약하고 있다. 특히, Johnson과 Wichern(2002, p. 708) 그리고 Izenman(2008, pp. 488-490)이 다차원척도법의 관점에서 새먼사상을 간단히 언급하고 있다. 이 절에서는 새먼의 비선형사상을 다차원척도법에 적용한 비선형 계량형 다차원척도법의 이론을 정리 요약하기로 하자. 새먼사상은 일반적으로 패턴인식에서 고차원 공간의 자료들 간의 관계를 잘 유지하도록 저차원 공간으로 비선형 사상시키는 매우 잘 알려진 기법으로 이에 대한 여러 가지 성질을 비교 분석한 Ridder와 Duin(1997)을 참고할 수 있다. 여기서 그들은 새먼사상 알고리즘의 단점은 새 자료 점이 사상되어야 할 때, 전체 사상과정이 반복되어야 하며 대신에 주성분분석과 같이 수리적이거나 대수적인 사상과정을 제공하지 못한다는 것이다. 이를 극복하는 다양한 방법이 제시되어 있고 사례들도 살펴볼 만하다.

7.1.1 비선형 모형

다변량자료행렬 X의 i번째 행벡터 $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ip})^t$는 관측된 i번째 개체를 나타내며 이들 n개의 행벡터 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$은 p차원의 공간에서 n개 개체들의 좌표점이다. 두 개체 $\boldsymbol{x}_i$와 $\boldsymbol{x}_j$ 사이의 거리를 d_{ij}라고 하고 이를 p차원의 공간에서 두 개체 간의 비유사성으로 고려하자. 대개 거리 d_{ij}는 유클리드거리가 사용된다. 따라서 일반적으로 다차원척도법이란 이 거리 d_{ij}와 차원축소된 $k(\le p)$차원의 공간에서 두 개체의 거리 δ_{ij} 사이의 관계가 서로 일치되도록 하는 것이다. 여기서 이들 간의 관계를 일반적으로 1.3.1절에서 임의의 함수 $f(\cdot)$에 대해서

$$d_{ij} = f(\delta_{ij}) + \epsilon_{ij},\ i,j = 1,\ldots,n$$

로 정의되며 여기서 ϵ_{ij}는 측정 및 $k(\le p)$차원 공간근사에 따른 왜곡오차인 오차항이다.

특별히 비선형 계량형 다차원척도법의 모형을 위해서 함수 $f(\cdot)$를 절대척도 모형인 $f(\delta_{ij}) = \delta_{ij}$로 두자. 그리고 새먼의 스트레스를

$$E \equiv \frac{1}{\sum_{i<j}^{n} d_{ij}} \sum_{i<j}^{n} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}} \tag{7.1}$$

로 정의하자. 이는 1.3.2절의 크루스칼이나 영의 스트레스 공식과 유사하나 식 (7.1)의 스트레스는 개체 간의 원 거리 d_{ij}에 의해서 동일하게 나누어져 가중화되어 있다. 이로 인하여 척도(scale)에 영향을 받기 쉽다. 특히, 새먼의 스트레스는 찾고자 하는 $k(\leq p)$차원의 형상공간에서 개체 i의 좌표점 $\boldsymbol{x}_i = (x_{i1}, ..., x_{ik})^t$가 k개 좌표 $x_{i1},...,x_{ik}$로 이루어져 있어 두 개체의 거리 δ_{ij}를 형성하는 nk개의 좌표점의 함수로 이루어져 있다. 그러므로 이 스트레스를 최소화하는 $k\ (\leq p)$차원 형상공간에서 크기가 $n \times k$인 형상좌표

$$X_{(k)} = (\boldsymbol{x}_1,...,\boldsymbol{x}_n)^t = (x_{ir}),\ i = 1,...,n;\ r = 1,...,k$$

를 찾는 것이다. 이는 지금까지 다차원척도법에서 얻고자 하는 형상공간의 형상좌표 개념과 동일하다.

7.1.2 비선형사상 알고리즘

식 (7.1)의 새먼의 스트레스는 찾고자 하는 $k(\leq p)$차원의 형상공간에서 두 개체의 거리 δ_{ij}를 형성하는 nk개의 좌표점의 함수로 이루어져 있다. 따라서 스트레스를 최소화하는 과정에서 최경사하강법과 같은 수치최적화 알고리즘이 활용되며 이를 [알고리즘 6: 새먼 비선형사상 알고리즘]으로 정리하였다. 이는 1.3.2절의 [알고리즘 1: 스트레스 최소화 알고리즘]에서도 언급한 바 있다.

참고로 [알고리즘 6]의 식 (7.3)에서 편도함수들은 다음과 같다.

$$\frac{\partial E}{\partial x_{ir}} = -\frac{2}{c} \sum_{j \neq i, j=1}^{n} \left[\frac{d_{ij} - \delta_{ij}}{\delta_{ij} d_{ij}} \right] (x_{ir} - x_{jr}),$$

$$\frac{\partial^2 E}{\partial x_{ir}^2} = -\frac{2}{c} \sum_{j \neq i, j=1}^{n} \frac{1}{d_{ij}\delta_{ij}} \left[(d_{ij} - \delta_{ij}) - \frac{(x_{ir} - x_{jr})^2}{\delta_{ij}} \left(1 + \frac{d_{ij} - \delta_{ij}}{\delta_{ij}} \right) \right].$$

Dzwinel(1994)에 따르면 식 (7.3)의 경험적으로 주어진 마요인(magic factor) α의 범위가 모든 문제에서 최적화가 될 것이라고 기대하기는 어렵다는 점을 지적한다. 대안으로 정규경사하강법(normal gradient descent)에 의해서 [알고리즘 6]의 [2단계]의 식 (7.3) 대신에 $t+1$번째에서 $k\ (\leq p)$차원 형상공간의 좌표로 다음을 사용할 수 있음을 제안하고 있다.

$$x_{ir}(t+1) = x_{ir}(t) - \frac{\partial E(t)}{\partial x_{ir}(t)}$$

[알고리즘 6] 새먼 비선형사상 알고리즘

1단계: t번째 반복 후 사상 오차

$$E(t) \equiv \frac{1}{c} \sum_{i<j}^{n} \frac{(d_{ij} - \delta_{ij}(t))^2}{d_{ij}} \tag{7.2}$$

를 정의한다. 여기서

$$c = \sum_{i<j}^{n} d_{ij} \text{이고 } \delta_{ij}(t) = \left[\sum_{r=1}^{k} (x_{ir}(t) - x_{jr}(t))^2 \right]^{1/2} \text{이다.}$$

2단계: $t+1$번째에서 $k\ (\leq p)$차원 형상공간의 좌표는

$$x_{ir}(t+1) = x_{ir}(t) - \alpha \frac{\partial E(t)}{\partial x_{ir}(t)} \Bigg/ \left| \frac{\partial^2 E(t)}{\partial x_{ir}(t)^2} \right| \tag{7.3}$$

이다. 여기서 $\alpha \approx 0.3$ 또는 0.4로 경험적으로 결정된 마요인이다.

일반적으로 비계량형과 베이즈 계량형 다차원척도법에서는 반복적으로 계산하는 수치해석적 알고리즘에서 초깃값으로 계량형 다차원척도법의 형상좌표를 사용한다. 이와 동일하게 비선형 계량형 다차원척도법의 비선형사상 알고리즘에서도 초깃값으로 이를 사용한다. 실제로 비선형 계량형 다차원척도법을 수행하기 위한 sammon 함수를 제공하는 R-프로그램에는 초깃값에 관한 인수로 계량형 다차원척도법을 위한 cmdscale 함수를 사용하고 있다.

7.2 비선형 계량형 다차원척도법의 예

이미 3.3절에는 10개의 아라비아 숫자의 모스부호(Morse code)에 대한 계량형과 비계량의 다차원척도법을 비교하기 위하여 SAS/MDS와 R-프로그램을 적용하였다. 이 절에서는 이 자료를 활용하여 비선형 계량형 다차원척도법을 실시하고 이를 위한 R-프로그램을 제공하여 얻어진 결과를 3.3.2절의 계량형 다차원척도법의 결과와 비교하려 한다.

7.2.1 모스부호 자료

7.1절에서 정리한 비선형 계량형 다차원척도법을 적용하기 위한 사례로 3.3.1절과 3.3.2절에서 사용한 Mardia 외 2인(1979, Chapter 14)의 점(. dot)과 선(- dash)의 짧은 신호로 구성된 모스부호를 고려하자. 여기서 0, 1, 2, 3, 4, 5, 6, 7, 8, 9로 이루어진 10개의 아라비아 숫자에 대한 〈표 3.2〉와 같은 유사성자료이다. 이 자료는 아라비아 숫자에 대한 유사성행렬 $C=(c_{ij}),\ i,j=1,\ldots,10$로 모스부호를 모르는 598명에게 청각적으로 기계에 의해서 두 수의 모스부호신호를 짝을 지어 두 번(1 다음 2, 2 다음 1과 같이) 들려 주고 두 신호의 짝이 같다고 생각하는 사람들의 비율을 나타낸다. R-프로그램에서 입력자료인 비유사성행렬 $D=(d_{ij}),\ i,j=1,\ldots,10$을 얻기 위해 이 유사성행렬부터 표준변환 $d_{ij}=\sqrt{c_{ii}-2c_{ij}+c_{jj}}$을 실시한다. 이 과정은 3.3.1절의 [프로그램 3.3]에 잘 나타나 있고 그 결과인 비유사성행렬자료는 [부록 2: 자료 10(파일이름: morse.txt)]에 정리되어 있다. 이를 입력자료로 하는 비선형 계량형 다차원척도법에 관한 R-프로그램이 [프로그램 7.1]이다.

7.2.2 R-프로그램과 결과

[프로그램 7.1]의 # Nonlinear Metric MDS 이하는 7.2.1절에서 설명한 모스부호자료를 사용하여 새먼의 비선형사상을 사용한 비선형 계량형 다차원척도법을 위한 R-프로그램이다. 자료를 불러들이는 형식은 이미 3.3.2절에서 설명한 바 있으니 참고하기를 바란다.

[프로그램 7.1] 모스부호의 비유사성자료의 비선형 계량형 다차원척도법을 위한 R-프로그램

```
# Nonlinear Metric MDS
d<-as.matrix(read.table("c://R-MDS//morse.txt", header=T, check.names=F))
d
```

```
library(MASS)
con<-sammon(d, k=2, magic=0.3)
con
x<-con$points[,1]
y<-con$points[,2]
lim<-c(-max(abs(con$points)), max(abs(con$points)))
plot(x,y, xlab="Dimension 1", ylab="Dimension 2", xlim=lim, ylim=lim)
text(x,y+0.6, colnames(d), cex=0.8)
abline(v=0, h=0)
```

[프로그램 7.1]을 보면 새먼의 비선형사상을 사용한 비선형 계량형 다차원척도법에 관한 sammon 함수를 이용하였고 그 이전에 library(MASS) 문에 의해서 라이브러리 MASS를 지정하였다. 함수에서 인수로 $k=2$는 2차원의 형상공간을 적합을 요구하며, 특히, sammon 함수에는 알고리즘에서 계량형 다차원척도법의 형상좌표를 초깃값으로 이용하기 위하여 인수로 y=cmdscale(d, k)에 의해 cmdscale 함수를 이용하는 것이 디폴트로 주어진다. 더불어 수치적 반복을 위해 niter=100과 수렴의 조건 tol=1e-4 역시 디폴트 인수로 주어진다. 이런 인수 설정은 3.3.2절의 비계량형 다차원척도법을 위한 isoMDS 함수와 동일하다. 다만 magic=0.3은 [알고리즘 6]의 [2단계]에서 정의한 마요인 α값에 관한 인수이다. 실제 이 값을 0.4로 하여도 결과는 거의 동일하다.

[프로그램 7.1]의 수행 결과는 [결과 7.1]에 정리되어 있다. 먼저 ①은 0(zero)-9(nine)로 이루어진 10개의 아라비아 숫자에 대한 비유사성행렬자료이다. ②의 $points에는 2차원의 형상좌표와 $stress에는 최종적합도 0.05가 주어져 있고 끝으로 ③ 비선형 계량형 다차원척도그림이 제공되고 있다.

[결과 7.1] [프로그램 7.1]의 수행결과

①

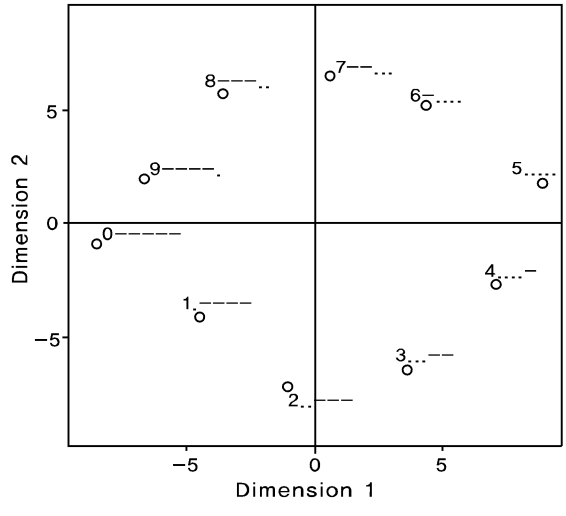
```
       1.---- 2..--- 3...-- 4....- 5..... 6-...., 7--... 8---.. 9----. 0-----
 [1,]  0.000   7.00  11.75 12.690 13.040  12.080 11.360  9.899  7.810  8.602
 [2,]  7.000   0.00   7.55 11.490 12.770  12.120 11.140 11.270 11.140 12.120
 [3,] 11.750   7.55   0.00  9.950 11.050  10.300 11.700 11.920 12.610 12.730
 [4,] 12.690  11.49   9.95  0.000  8.185  10.340 11.220 12.290 12.880 13.000
 [5,] 13.040  12.77  11.05  8.185  0.000  10.770 11.790 12.570 13.080 13.190
 [6,] 12.080  12.12  10.30 10.340 10.770   0.000  6.403 11.400 12.690 12.000
 [7,] 11.360  11.14  11.70 11.220 11.790   6.403  0.000  6.557 10.680 12.210
 [8,]  9.899  11.27  11.92 12.290 12.570  11.400  6.557  0.000  7.937 10.200
 [9,]  7.810  11.14  12.61 12.880 13.080  12.690 10.680  7.937  0.000  5.196
[10,]  8.602  12.12  12.73 13.000 13.190  12.000 12.210 10.200  5.196  0.000
```

②

```
$points
           [,1]       [,2]
 [1,] -4.6918657 -4.1479500
 [2,] -1.1067790 -7.0914092
 [3,]  3.6649881 -6.3675462
 [4,]  7.2302568 -2.6164172
 [5,]  8.9676205  1.7923218
 [6,]  4.3613078  5.0984106
 [7,]  0.5899136  6.5070161
 [8,] -3.6744923  5.7627283
 [9,] -6.7091409  1.9101689
[10,] -8.6318088 -0.8473231
```

```
$stress
[1] 0.04812271
```

③



7.2.3 결과의 해석

[결과 7.1]을 보면 ②의 최종적합도 0.05는 2차원의 형상좌표에 의한 ③ 비선형 계량형 다차원척도그림이 〈표 1.9〉의 크루스칼 판별기준에 따르면 〈0.05 좋음(good)〉에 해당한다. 이 다차원척도그림은 전반적으로 1축의 왼편 아래쪽부터 오른편으로 0(-----), 1(.----), 2(..---), 3(...--), 4(....-), 5(.....), 6(-....), 7(--...), 8(---..), 9(----.) 순서로 매끈한 원을 형성하고 있다. 즉, 수평축(1차원축)은 왼쪽에서 오른쪽으로 갈수록 모스부호인 점(.)의 증가되는 추세를 측정하며 그 반면에 수직축(2차원축)은 아래쪽은 아라비아 숫자 1(.----), 2(..---), 3(...--), 4(....-)로 점(.)이 먼저 시작되며 위쪽은 6(-....), 7(--...), 8(---..), 9(----.)로 선(-)이 먼저 시작되어 신호의 동질성을 측정한다. 신호의 동질성을 나타내는 2차원축에 대하여 그 경향성이 벗어난 5(.....)와 0(-----)을 제대로 분류하기 위해서는 전체적으로 축의 회전을 반 시계 방향으로 30도 정도 할 필요가 있다.

7.3 계량형과 비선형 계량형 다차원척도법의 비교

지금까지 일반적으로 다차원척도법은 고차원의 두 개체 간의 비유사성으로 정의한 거리 d_{ij}와 차원 축소된 $k(\leq p)$ 차원의 공간에서 두 개체 사이의 거리 δ_{ij}사이의 관계가 서로 일치되도록 하는 것이 기본 개념이다. 계량형과 비선형 계량형 다차원척도법도 더불어 두 방법 모두 다음과 같은 기본모형을 가정한다.

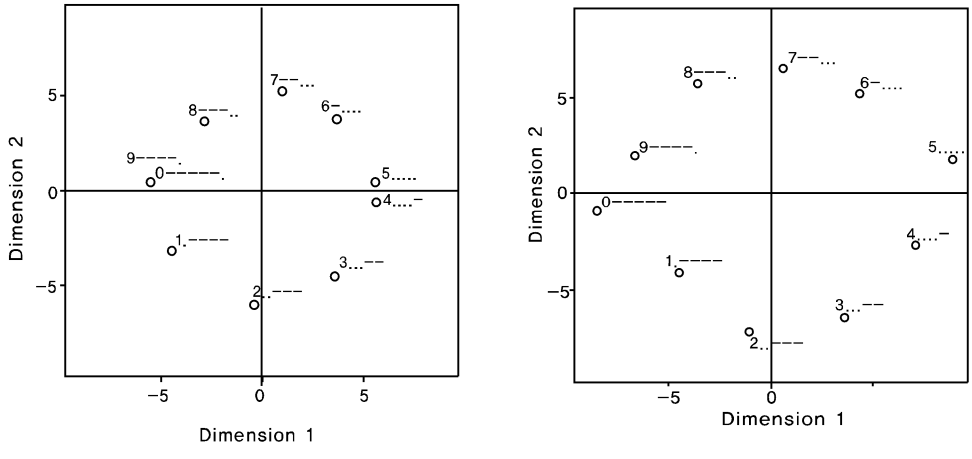
$$d_{ij} = \delta_{ij} + \epsilon_{ij},\ i,j = 1,...,n.$$

여기서 ϵ_{ij}는 측정 및 $k(\leq p)$ 차원 공간근사에 따른 왜곡오차인 오차항이다. 이 절에서 두 방법의 비교를 위해서 R-프로그램에서 제공하는 계량형 다차원척도법과 비선형 계량형 다차원척도법의 결과를 비교하고자 한다. R-프로그램에서 비선형 계량형 다차원척도법을 위한 sammon 함수를 사용할 때 반복적인 수치적 알고리즘의 초깃값으로 계량형 다차원척도법을 위한 mdscale 함수를 활용한다는 점을 이미 7.2.2절에서 설명한 바 있다. 다만 비선형 계량형 다차원척도법은 새먼의 스트레스 공식 식 (7.1)을 최소화시키는 [알고리즘 6: 새먼 비선형사상 알고리즘]을 활용하고 계량형 다차원척도법은 [알고리즘 2: 토거선 알고리즘]을 활용하여 $k\ (\leq p)$차원 형상공간에서 크기가 $n \times k$인 형상좌표 $X_{(k)}$ 를 찾는 것이다. 더불어 계량형 다차원척도법의 다차원척도그림이 주성분분석의 대수적 성질과 기하적 성질을 따르므로 최용석과 신상민(2013, 1.6절)의 주성분행렬도 관점에서 기하적인 해석이 유용하게 사용됨을 3.3.1절에서 이미 설명한 바 있다. 반면에 비선형 계량형 다차원척도법은 비록 알고리즘의 초깃값으로 계량형 다차원척도법의 형상좌표를 사용하지만 계량형 다차원척도법의 이러한 기하적 성질이 유지된다고 말할 수 없다.

두 방법의 실제 사례를 활용한 비교를 위해 ① 형상좌표 ② 최종적합도 ③ 다차원척도그림은 계량형 다차원척도법의 경우 3.3.2절의 [결과 3.5]를 참고하여 재인용하였고, 비선형 계량형 다차원척도법은 [프로그램 7.1]의 수행 결과인 [결과 7.1]을 참고하였다. 특별히 〈그림 7.1〉에는 이들의 다차원척도그림만 재인용하였다.

〈그림 7.1〉의 (a) 계량형 다차원척도그림과 (b) 비선형 계량형 다차원척도그림의 최종적합도는 각각 0.56과 0.05로 계량형 다차원척도그림보다는 비선형 계량형 다차원척도그림이 〈표 1.9〉 크루스칼의 판별기준에 따라 상대적으로 적합도가 좋다. 전반적으로 다차원척도그림에 대한 해석은 3.3.2절과 7.2.3절에 자세히 언급하고 있다. 다만 일부 모스부호가 1축의 왼쪽으로부터 다소 겹친 원을 형성한 (a) 계량형 다차원척도그림보다는 (b) 비선형 계

량형 다차원척도그림이 보다 모스부호가 고루 분포된 매끄러우면서 큰 원을 형성하고 있다. 즉, 계량형 다차원척도그림 (a)에서 좌표점을 나타내는 숫자를 원점에서 시작되는 벡터로 보면 0과 9, 그리고 4와 5는 각각 그들이 서로 이루는 각이 매우 작아 매우 연관성이 높게 여겨진다. 반면에 비선형 계량형 다차원척도그림 (b)에서는 다소 이들의 기하적 성질이 줄어들고 있다. 그러나 실제로 비선형 계량형 다차원척도법에서는 비선형사상 알고리즘의 초깃값만 계량형 다차원척도법을 사용하고 있기 때문에 이러한 기하적 성질이 유지된다고 말할 수 없다.



〈그림 7.1〉 (a) 계량형 다차원척도그림 (b) 비선형 계량형 다차원척도그림

연습문제

7.1 [연습문제 6.4]에는 로이드은행의 1905 ~ 1909년과 1925 ~ 1929년 사이에 각각 80명의 은행 종사원의 고용 기록에 대한 자료를 활용하여 얻어진 80×80 비유클리드거리로 된 비유사사성행렬 자료를 소개하고 있다. 이들은 각각 [부록 2: 자료 3(파일이름: Lloyds05.sd2, Lloyds05.xlsx), 자료 4(파일이름: Lloyds25.sd2, Lloyds25.xlsx]에 SAS와 엑셀자료세트로 각각 제공되어 있다.

1) 비선형 계량형 다차원척도법을 실시하고 해석하라.

2) 적절한 차원수를 정하고 적합도에 대하여 해석하라.
3) [연습문제 6.4]의 계량형과 베이즈 다차원척도법의 결과와 비교하라.

7.2 [연습문제 6.2]에서 Mardia 외 2인(1979, p. 238)에서는 6기 무덤에서 소장된 다섯 종류의 도자기 형태의 유사성에 관한 고고학 자료를 소개하고 있다. 어떤 무덤이 비슷한 종류의 도자기를 소장하고 있는지를 비유사성행렬을 통하여 비선형 계량형 다차원척도법을 실시하고 [연습문제 6.2]의 4)의 계량형과 베이즈 계량형 다차원척도법과의 결과를 비교하라.

7.3 6.2.1절에는 유럽 10개 도시(London-Dublin) 간의 거리에 대한 2차원 계량형과 베이즈 계량형 다차원척도법의 적합값과 다차원척도그림을 해석하고 비교하고 있다. 추가로 비선형 계량형 다차원척도법을 실시하여 서로 비교하라.

7.4 [연습문제 4.2]의 네 가지 식용작물(감자, 시금치, 상치, 튜나)에 대한 비유사성의 순위를 매겨 얻은 두 사람의 자료가 있다.

1) 평균에 의한 비유사성행렬 자료를 구하라.
2) 계량형 다차원척도법을 실시하고 해석하라.
3) 베이즈와 비선형 계량형 다차원척도법을 실시하고 비교하라.

7.5 [연습문제 3.4]에는 미국 내 22개 공익회사에 대한 평가를 위해 8개 변수(X_1: 고정요금 보상율, X_2: 자금 수익률, X_3: KW당 비용, X_4: 연부하율, X_5: 1974 ~ 1975년 사이 절정의 kWh 수요성장, X_6: 판매량, X_7: 핵비율, X_8: 총연료비용)로 측정한 자료가 주어져 있다.

1) 비유사성행렬로 피어슨거리와 마할라노비스거리를 구하여 보라.
2) 이들 거리의 비선형 계량형 다차원척도법 위한 적절한 차원수와 적합도를 구하라.
3) 2)에서 구한 각각의 거리에 대한 다차원척도그림을 통해 적절한 군집과 특성을 해석하고 비교하라.
4) 계량형 다차원척도법과의 결과를 비교하라.

부록 1
SAS MDS 절차

SAS/STAT에서 다차원척도법을 제공하는 MDS 절차가 6.07판 이후 현재 9.2판인 SAS Institute Inc.(2008, Chapter 5)에 나와 있다. 더불어 9.2판 이후에 ODS GRAPHICS ON 문은 다차원척도법과 관련된 그림(다차원척도그림, 스트레스-차원수 그림, 상그림, 세퍼드 그림 등)을 자동으로 쉽게 제공한다. 이미 6장 베이즈 다차원척도법을 제외하고 5장까지 다양한 사례들에 대한 전형적인 이원과 삼원, 최대우도 다차원척도법을 위한 PROC MDS 프로그램을 사용하였다.

[부록 1.1]에서는 PROC MDS 문법의 전반에 대해서 설명하고 [부록 1.2]에서는 9.2판 이후에 그림을 용이하게 얻게 하는 ODS GRAPHICS에서 MDS와 관련된 것을 설명하려 한다. 4장의 삼원과 5장의 최대우도 다차원척도법을 위한 범용 PROC ALSCAL과 PROC MLSCALE을 수용한 PC-윈도우용 PROC MDS와의 비교를 [부록 1.3]과 [부록 1.4]에서 각각 요약하기로 하자.

〈표 1〉 PROC MDS 절차

```
1) PROC MDS <options>;
2)          VAR variables;
3)          INVAR variables;
4)          ID/OBJECT variable;
5)          MATRIX/SUBJECT variable;
6)          WEIGHT variables;
7)          BY variables;
```

〈표 1〉의 문법형식에서 반드시 필요한 부분은 1)의 PROC MDS 문이다. 2)의 VAR 문의 variables는 입력 자료의 각 개체를 나타내는 변수를 지정하나 일반적으로 이 VAR 문의 활용도는 낮다. 3)의 INVAR 문의 variables는 옵션 〈INITIAL =SASdataset〉에서 지정된 SAS 자료세트에 들어 있는 다차원척도그림을 위한 형상좌표를 제공하는 변수명을 지정하여 초깃값으로 사용한다. 실제 이 문을 생략한다면 변수명 DIM1～DIMm(여기서 m은 최대 차원의 수)이 디폴트로 사용된다. 4)의 ID(/OBJECT) 문은 각 개체의 라벨을 위한 변수를 지정한다. 그러나 이 ID 문은 1.4절의 [프로그램 1.1]-[프로그램 1.3]과 같은 전형적인 형식을 이용한다면 없어도 된다. 5) MATRIX/SUBJECT 문은 주체(subject)에 대하여 개체들의 비유사성행렬이 존재하는 경우에 대한 4장의 삼원과 5장의 최대우도 다차원척도법에서 주체와 관련된 변수를 지정하게 한다. 6) WEIGHT 문은 비유사성을 측정할 때 고려되는 가중치(weight)를 포함하는 수치형변수를 지정한다. 이 문에서 지정될 변수의 개수와 순서는 입력자료의 각 개체를 나타내는 변수와 일치한다. 그리고 이 문이 지정되지 않았다면 가중치는 동일하게 지정된다. 7) BY 문은 SAS의 여러 절차에 두루 사용되므로 여기서 설명은 생략한다.

지금까지 〈표 1〉의 PROC MDS 절차의 일반형식을 간략하게 소개하였다. 이 중에서 일반적으로 SAS의 여러 절차에서 두루 사용되는 전형적인 2) VAR 문, 4) ID 문 그리고 7) BY 문에 대한 자세한 설명과 활용은 생략하고 다차원척도법에 관련된 PROC MDS 문법을 본격적으로 설명하려고 한다.

부록 1.1 PROC MDS 문의 옵션

PROC MDS 문의 〈options〉에는 크게 49가지의 옵션을 사용할 수 있다. 이 옵션을 사용하지 않은 디폴트의 PROC MDS는 단지 스트레스 값을 찾는 수치적 반복계산과정만을 제공한다. 9.2판인 SAS Institute Inc.(2008, Chapter 5) 이후에는 PROC MDS 문을 실행하기 전에 ODS GRAPHICS ON 문을 실행한다면 다차원척도법과 관련된 그림(다차원척도그림, 스트레스-차원수 그림, 상그림, 세퍼드그림 등)을 자동으로 쉽게 얻게 된다. 이는 최용석(1995, 4장)에서 참고한 SAS Institute Inc.(1985)의 6.07판과 같은 구판에서 PROC PLOT 문을 이용하는 수고를 덜어 주었다. ODS GRAPHICS ON 문에서 제공하는 그림에 대한 출력결과를 PROC TEMPLATE 문에서 수정이 가능하며 여기에 대해서는 [부록 1.2]에서 정리하고 있다. 〈표 2〉에는 PROC MDS 문의 옵션을 기능별로 나누어 정리하였고 이에 대하여 상세히 설명하고자 한다.

〈표 2〉 PROC MDS 문의 옵션

측정수준 지정 (MDS 모형의 종류)	LEVEL=
자료의 형태	SIMILAR=, COEF=, CONDITION=, SHAPE=
자료세트의 지정	DATA=, OUT=, OUTFIT=, OUTRES=, INITIAL=
스트레스 최적척도화	DIMENSION= 〈to〉, FIT=, FORMULA=, ALTERNATE=
OUT=dataset에 저장	OCOEF, OCONFIG, OCRIT, OTRANS, OUTITER
출력결과 조정	NOPHIST, PCOEF, PCONFIG, PDATA, PFINAL PFIT, PFITROW, PINAVDATA, PINEIGVAL PININ, PINIT, PITER, PTRANS
결측치 조정	CUTOFF=, INAV=, NOULB
수렴문제 조정	CONVERGE=, GCONVERGE=, MCONVERGE= MAXITER=, MINCRIT=, OVER=, RIDGE=
기타 선택	DECIMALS=, EPSILON=, NEGATIVE, NONORM RANDOM=, SINGULAR=, UNTIE

부록 1.1.1 측정수준 지정

LEVEL=ABSOLUTE/INTERVAL/RATIO/ORDINAL/LOGINTERVAL (=ABS/INT/RAT/ORD/LOG =A/I/R/O/L)	디폴트는 ORDINAL이다.

이들 측정수준에 대한 설명은 1.3.1절에서 하였다. 계량형 다차원척도법을 위해서는 〈LEVEL= ABSOLUTE/INTERVAL/RATIO〉를 사용하고, 비계량형 다차원척도법을 위해서는 〈LEVEL= ORDINAL〉(디폴트)이 사용되며 5장과 6장을 제외한 여러 장에 걸쳐 활용되고 있다. 특히 〈LEVEL= LOGINTERVAL〉은 5장의 최대우도 다차원척도법에서 사용한다. 범용 PROC ALSCAL에서는 〈LEVEL=NORMINAL/INTERVAL/RATIO/ORDINAL〉 옵션이 사용되었으나 PROC MDS에서는 〈NORMINAL〉이 빠지고 〈ABSOLUTE〉가 대신 지원되며 [부록 1.3]의 〈표 3〉에서 이들을 서로 비교하고 있다.

부록 1.1.2 자료의 형태

SIMILAR SIMILAR=max (SIM=max)	입력자료가 유사성행렬일 경우 이를 비유사성행렬로 바꾸라는 옵션이다.
COEF=IDENTITY/DIAGONAL (COEF=IDEN/DIAG =I/D)	디폴트는 IDENTITY로 이원 다차원척도법에 해당한다. 삼원 다차원척도법에서는 DIAGONAL을 사용한다.
CONDITION=UN/MATRIX/SUBJECT/ROW (CON=U/M/S/R)	입력자료의 조건을 지정하는 옵션이다. 〈SHAPE=〉 옵션과 같이 사용되며 디폴트는 MATRIX이다.
SHAPE=TRIANGULAR/SQUARE (=TRI/SQU=T/S)	입력자료의 행렬형태를 지정하는 옵션이다. 〈CONDITION=〉 옵션과 같이 사용된다.

유사성행렬 $C=(c_{ij}), i,j=1,..n$이 프로그램에서 입력자료로 사용될 경우 〈SIMILAR〉 옵션을 사용한다. 특히, 옵션 〈SIMILAR=max〉에서 =max를 지정하지 않으면 유사성행렬에서 대각원소를 포함하여 가장 큰 유사성 $Max=Max(c_{ij})$를 찾고, 만약에 =max를 지정하면 max값과 유사성행렬에서 대각원소를 포함하여 가장 큰 유사성 $Max=Max\{(c_{ij}), \max\}$을 찾아 가장 단순한 변환 $d_{ij}=Max-c_{ij}$을 통해 얻어지는 비유사성행렬 $D=(d_{ij})$, $i,j=1,..n$을 가지고 분석하게 된다. 그리고 대각원소에 ·(missing value)나 0(zero)이 있는 경우 〈SIMILAR=max〉 옵션을 꼭 사용하고 대각원소에 가장 큰 유사성이 있는 경우 디폴트로 이 옵션을 사용하지 않아도 된다. 이는 이미 1.1.3절의 [사례 3]에서 대각원소에 가장 큰 유사성 $Max=9$가 있는 유사성행렬이 제공되어 있어 1.4.3절의 [프로그램 1.3]에서 디폴트로 처리되어 있다.

특히 〈COEF=DIAGONAL〉은 삼원 다차원척도법을 위한 INDSCAL 모형을 지정하는 것으로 결과에서 각 주체별 가중치를 제공하는 차원계수(dimension coefficients)를 제공하며 활용의 예는 4.2.2절의 [프로그램 4.2]에서 참고하면 된다. 4.1.2절에서 이 차원계수의 기하적 역할에 대해서 자세히 설명하고 있다.

〈CONDITION=〉과 〈SHAPE=〉 이들 두 옵션은 입력자료의 조건과 행렬형태에 따라 다르게 사용된다. 입력자료의 조건에 대해서는 Young(1987, pp. 60-63)은 5가지를 정의하고 있다. 특히, PROC MDS에서 행렬-조건(matrix-conditional), 조건 없이(unconditional), 행-조건(row-conditional)을 채택했고, 열-조건(column-conditional)과 일반-조건(general-conditional)이 나머지에 입력자료 조건에 해당한다.

행렬형태는 대칭이면서 삼각행렬형태인 비유사성행렬 하나가 분석에서 이용되는 3장의

이원 다차원척도법과 각 주체에 따라 여러 개의 비유사성행렬이 분석에 사용되는 5장의 최대우도 다차원척도법이 있다. 이 경우에 〈CONDITION=MATRIX SHAPE=TRIANGULAR〉를 활용하면 된다. 특히 〈SHAPE=TRIANGULAR〉는 〈CONDITION=MATRIX〉 옵션에 대해 디폴트이고 실제로 〈CONDITION=MATRIX〉도 디폴트이다. 그리고 4장의 삼원 다차원척도법에서처럼 각 주체에 대해 비대칭이고 정방행렬 형태의 입력자료인 경우 〈CONDITION=ROW SHAPE=SQUARE〉 옵션을 사용한다. 이때는 〈SHAPE=SQUARE〉 옵션이 〈CONDITION=ROW〉 옵션에 대해 디폴트이다. 끝으로 그룹별로 나뉘어진 행렬자료(split-file data)인 경우는 〈CONDITION=UN SHAPE=SQUARE〉 옵션을 사용한다.

부록 1.1.3 자료세트의 지정

DATA=SASdataset	PROC MDS에 의해서 분석될 자료세트의 이름을 지정한다. 이는 SAS프로그램의 전형적인 옵션이다.
OUT=SASdataset	다차원척도그림을 위한 형상좌표에 관련된 모든 결과를 포함하는 새로운 자료세트를 생성한다. 만약에 〈OCOEF, OCONFIG, OCRIT, OTRANS, OUTITER〉 옵션 중 하나 혹은 그 이상을 지정한다면 지정한 옵션에 관련된 결과만 SASdataset에 저장된다.
OUTFIT=SASdataset	부적합도(badness-of-fit criterion)를 포함하여 적합도에 관련된 척도를 저장한다.
OUTRES=SASdataset	MDS 모형이 추정되고 일반적 회귀분석에서처럼 잔차들이 저장된다. 따라서 MDS 모형의 적합성을 보여 주는 세퍼드그림 등을 제공받기 위한 옵션이다.
INITIAL=SASdataset	MDS 모형의 모수들에 대한 초깃값을 설정하는 자료세트를 지정한다. 디폴트는 자료로부터 초깃값이 계산된다.

9.2판 이전에는 다차원척도법과 관련된 다차원척도그림, 스트레스-차원수 그림, 상그림, 세퍼드그림 등을 얻기 위해서는 PROC MDS 문에 두 옵션 〈OUT=SASdataset〉과 〈OUTRES=SASdataset〉을 지정하고 PROC PLOT 문에서 이들 자료세트를 이용하고는 하였다. 그러나 9.2판에서는 세퍼드그림을 제외하고는 ODS GRAPHICS ON 문을 PROC MDS 문 이전에 실행하면 자동으로 얻게 되어 이들 두 옵션에 대한 활용도가 낮아졌다. 그러나 [프로그램 3.2]와 [결과 3.2]에서와 같이 스트레스 계산을 포함하여 모형의 적합과정을 수치적으로 확인하고 이해하는 작업을 위해서는 여전히 필요하다.

〈OUT=SASdataset〉 옵션에서 자료세트로 저장하고자 하는 결과에 관련된 옵션 〈OCOEF, OCONFIG, OCRIT, OTRANS, OUTITER〉에 대하여 다음과 같이 요약할 수 있다.

OUT=SASdataset

OCOEF	차원계수(dimension coefficient). 이 옵션의 활용은 4장의 삼원 다차원척도법에서 찾아볼 수 있다.
OCONFIG	형상좌표(configuration coordinates).
OCRIT	부적합도. 이 옵션은 스트레스-차원수 그림(1.3.3절을 참고)을 얻을 때 유용하다.
OTRANS	변환모수의 추정치. 비계량형 다차원척도법을 제공하는 〈LEVEL=ORDINAL(디폴트)〉이 지정되어 있다면 이 옵션에 해당하는 결과는 없다. 그 이유는 비계량형 다차원척도법은 모수적 표현이 없는 단조변환에 의한 순서척도모형이기 때문이다.
OUTITER	스트레스 최소화 알고리즘의 수치적 계산과정에서 각 반복마다 형상좌표와 부적합도를 제공한다. 이 옵션은 특별히 〈OUTFIT=〉과 〈OUTRES=〉에서 지정된 자료세트에도 결과를 저장한다.

부록 1.1.4 스트레스 및 최적척도화

DIMENSION=n 〈to m〉 (DIMENS/DIM=)	디폴트는 〈DIMENSION=2〉로 다차원척도법에서 차원의 수를 정하여 준다. 차원수 결정을 위해서는〈DIMENSION=1 to 4〉와 같이 1차원부터 4차원까지의 다차원척도법의 결과와 스트레스-차원수 그림을 제공받을 수 있다. 여기서 n과 m은 1보다 커거나 같아야 하고 개체의 수보다 작아야 한다.
FIT=DISTANCE/SQUARED/LOG (=DIS/SQU/LOG=D/S/L=1/2/0)	디폴트는 〈FIT=DISTANCE〉이다. 스트레스 공식과 관련되어 d_{ij}와 δ_{ij}의 관계를 주는 최적변환을 지정하는 옵션이다.
FORMULA=0/1/2 (FOR=OLS/USS/CSS=O/U/C)	디폴트는 〈FORMULA=1〉이다. 〈FIT=LOG〉는 최대우도 다차원척도법의 옵션으로 〈FORMULA=2〉가 디폴트이다. 옵션 〈FORMULA=0〉은 각 스트레스공식에서 수정척도인자(SC)가 없는 형태를 의미한다.
ALTERNATE =NONE/MATRIX/SUBJECT/ROW〈=n〉 (ALT=N/M/S/R 〈=n〉)	교대최소제곱법 형태의 결정을 위한 옵션이다.

이 옵션들은 1.3.2절의 스트레스 공식과 관련되어 있고 비유사성 d_{ij}와 단조함수 $f(\delta_{ij})$의 입력형태에 대한 것이다. 특히, 디폴트 〈FIT=DISTANCE〉 옵션은 식 (1.7)과 식 (1.8)에서 비유사성 d_{ij}와 단조함수 $f(\delta_{ij})$에 의한 최적의 $f(\cdot)$을 찾기 위한 크루스칼의 스트레스 1과 2인 $ST1$ 과 $ST2$를 이용한다. 〈FIT=SQUARED〉 옵션은 식 (1.9)와 식 (1.10)에서 d_{ij}^2와 $f(\delta_{ij})^2$에 의한 최적의 $f(\cdot)$를 찾게 하는 Young의 S-스트레스 1과 2인 $SST1$과 $SST2$를 이용하게 한다. 끝으로 〈FIT=LOG〉 옵션은 각 주체 $s=1,...,m$에 대하여 $\log d_{ij}^{(s)}$와 $\log f(\delta_{ij}^{(s)})$를 이용하게 하며 이는 최대우도 다차원척도법과 관련된 옵션으로 5장을 참고하기 바란다.

〈FORMULA=0/1/2〉 옵션들은 앞서 설명한 〈FIT=〉 옵션과 함께 사용되며 이를 스트레스의 표준화형태에 관한 옵션이라고 한다. 표준화는 수정척도인자 식 (1.6)에 의해서 이루어지며 스트레스 공식에서 수정척도인자가 없다면 〈FORMULA=0〉을 지정하며 특히 〈LEVEL=ORDINAL〉 옵션과 사용할 수 없다. 식 (1.8)과 식 (1.10)과 같이 수정척도인자가 총평균에 대해 수정되었다면 〈FORMULA=2〉 옵션을, 수정되지 않았다면 〈FORMULA=1〉 옵션을 지정한다. 다시 말해서 크루스칼의 스트레스 1과 2인 $ST1$과 $ST2$의 차이는 수정척도인자가 총평균에 대하여 수정되었는지에 따라 정해지며 Young의 S-스트레스 1과 2인 $SST1$과 $SST2$의 차이도 동일하다. 〈FORMULA=1〉 옵션은 〈FIT=LOG〉 옵션과는 사용될 수 없다.

참고로 〈LEVEL=ORDINAL(디폴트) FIT=DISTANCE FORMULA=1〉 옵션은 크루스칼의 $ST1$에 해당하며, 〈LEVEL=ORDINAL(디폴트) FIT=DISTANCE FORMULA=2〉 옵션은 $ST2$에 해당한다. 〈LEVEL=ORDINAL(디폴트) FIT=SQUARED FORMULA=1〉 옵션은 Young의 $SST1$에 해당하며, 〈LEVEL=ORDINAL(디폴트) FIT=SQUARED FORMULA=2〉 옵션은 $SST2$에 해당한다. 그리고 [결과 1.1]-[결과 1.3]과 같이 PROC MDS 프로그램 수행결과의 제일 첫 머리에 어떤 스트레스 공식을 사용하였는지가 이들 옵션과 관련하여 제공된다. 이들에 대한 활용과 관련된 스트레스 공식은 1.3.2절을 참고 바란다.

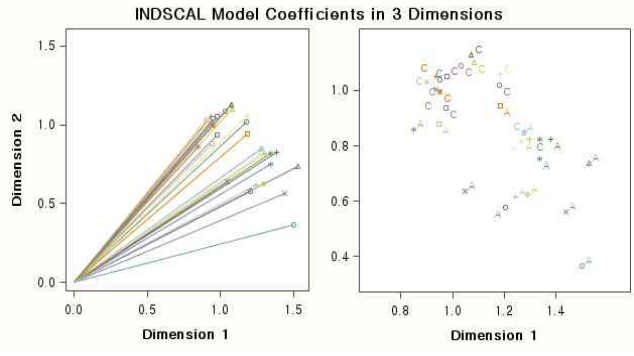
〈ALTERNATE=NONE/MATRIX/SUBJECT/ROW〈=n〉〉 옵션은 MDS 모형에서 최적화시킬 $f(\cdot)$를 찾을때 최소제곱법에 의한 모형추정단계와 자료 최적척도화 단계를 교대로 반복하는 ALSOS 알고리즘을 지정한다. 이 알고리즘은 SAS의 ALSCAL절차에서 구현된 교대최소제곱(ALS, alternating least squares)방법에 의한 최적척도화(OS, optimal scaling) 기법이다. 이들의 수치적 연산은 메모리를 많이 요구한다. 이들 옵션(NONE/MATRIX/ROW)은 메모리를 덜 사용하는 순서로 되어 있다. 일반적으로 디폴트로 하여도 무난하다고 생각

된다. 여기에 대한 자세한 설명은 Young(1987, pp. 199-203)과 허명회(1994, p. 4)에 잘 나와 있다.

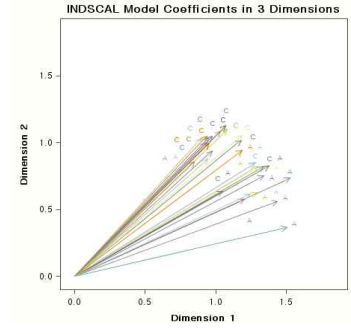
부록 1.1.5 출력결과 조정

NOPHIST(NOP)	스트레스 최소화알고리즘의 수치적 계산과정이 출력되지 않는다.
PCOEF	추정된 가중치에 관한 차원계수를 출력한다.
PCONFIG	형상좌표를 출력한다.
PDATA	자료행렬을 출력한다.
PFINAL	스트레스 최소화 알고리즘의 수치적 계산과정의 마지막 부적합도를 출력한다. 일반적으로 〈NOPHIST PFINAL〉 옵션을 같이 사용하는 것이 바람직하다.
PFIT	PFINAL 결과와 대동소이하다.
PFITROW	행 부적합도를 출력한다. 이 옵션은 4장의 삼원 다차원척도법과 관련 된 것으로〈CONDITION=ROW〉 옵션과 함께 사용된다.
PINAVDATA	옵션 〈INVA=DATA(디폴트)〉에 의해서 초기화동안에 계산되는 자료행렬의 가중평균과 가중합을 출력한다.
PINEIGVAL	초기 형상좌표를 제공할 행렬의 고유값을 출력한다.
PININ	〈INITIAL=dataset〉의 초기 형상좌표.
PINIT	초기 부적합도와 초기 형상좌표 및 변환모수의 초깃값.
PITER	스트레스 최소화 알고리즘의 수치적계산과정의 각 반복에서 부적합도, 형상좌표 및 변환모수 값을 출력한다.
PLOTS〈(flip)〉 =COEFFICIENT (ONE)/NONE〉	(flip)은 다차원척도그림과 차원계수그림의 차원축을 서로 바꾸어 출력한다. 〈=COEFFICIENT(ONE)〉은 삼원 다차원척도법에서 차원계수그림을 하나로 합친 형태로 출력한다. 〈=NONE〉 모든 그림을 출력하지 않는다.
PTRANS	변환모수를 출력한다.

다음의 결과는 [부록 1.1.5] 출력결과 조정 옵션 중에서 삼원 다차원척도법의 출력결과를 조정하는 옵션 〈PLOTS=COEFFICIENTS(ONE)〉에 대한 활용의 예이다. [프로그램 4.2]의 [결과 4.2]에서 차원별 각 주체에 대한 차원계수에 대한 그림으로 이 옵션을 PROC MDS 문에 (a) 사용하지 않은 경우 (b) 사용한 경우이다. 그 차이점은 (b)의 경우 (a)의 두 그림을 결합한 형태이다.



(a) 사용하지 않은 경우



(b) 사용한 경우

부록 1.1.6 수렴문제 조정

CONVERGE=p (CONV=)	스트레스 최적화알고리즘의 단조(monotone)와 기울기(gradient) 두 종류의 수렴조건을 지정한다. $0 \le p \le 1$이고 디폴트는 p=0.01이다. 0.0001보다 작으면 수렴이 만족되지 않는다.
GCONVERGE=p (GCONV=)	기울기 수렴조건만 지정한다.
MCONVERGE=p (MCONV=)	단조 수렴조건만 지정한다.
MAXITER=n	n(≥0)은 최대 반복수를 나타낸다. 디폴트는 n=100이다.
MINCRIT=n (CRITMIN=)	부적합도기준이 n(≥0)이하이면 스트레스 최소화 알고리즘의 수치적 계산이 멈춘다. 디폴트는 n=1E-6(=10^{-6})이다.
OVER=n	n≥1일 경우 최대과이완인자(maximum overrelation factor)를 지정하며 수렴되지 않을 때 사용 한다. 일반적으로 n이 1과 2사이에 존재하고 디폴트 n=2는 〈LEVEL=ORDINAL ALTERNATE=MATRIX 또는 ALTERNATE=ROW〉일 때 이다. 그 외에 디폴트는 n=1이다.
RIDGE=n	수치적 연산에서 부동소수점 오버플로우(floating point overflow)를 방지하기 위한 초기 능형(ridge)값 n(≥0)을 지정한다. 만약에 이런 현상이 일어났다면 n=0.01, 1 또는 100처럼 크게 지정한다. 디폴트는 n=1E-4. 좋은 초기 추정값이 있다면 n=0이 수렴을 빠르게 한다.

스트레스 최소화 알고리즘을 위한 수렴방법으로 특히 반복계산의 수렴문제에 대해서는

Kennedy와 Gentle(1980, Chapter 10)을 참고하기를 바란다. 컴퓨터 내부에서 정수는 고정소수점(fixed point)으로 실수는 부동소수점(floating point)으로 표현된다. 수치적 연산 중 최대정수보다 더 큰 정수가 일어나는 현상을 오버플로우(overflow)라고 하며 이때는 컴퓨터가 실행을 중지한다(Conte와 de Boor, 1980, pp. 7-11). 이 경우에 〈OVER=n〉 옵션을 사용하여 1과 2 사이에 존재하는 n을 지정하면 된다.

부록 1.1.7 결측치 조정

CUTOFF=n	디폴트는 CUTOFF=0이다. n보다 작은 자료는 결측치로 처리하라.
INAV=DATA/SSCP (=D/S)	초기좌표계산에서 결측치를 추정하여 형상좌표를 제공하는 행렬을 계산한다. 이 경우 〈=DATA〉는 자료행렬의 가중평균을 계산하고 〈=SSCP〉는 가중평균이 없는 형상좌표를 제공하는 행렬로 만들라는 의미이다. 디폴트는 〈=DATA〉이다.
NOULB	디폴트이면 결측치는 Rabinowitz 상하한보간법에 의해서 추정된다. 이 옵션은 이를 방지한다.

부록 1.1.8 기타 선택

DECIMALS=n (DEC=)	소수점 이하의 자리수를 지정하며 디폴트는 n=2이다.
EPSILON=n (EPS=)	n(0〈n〈1)은 MDS모형에서 제곱거리에 더해진 양으로 0으로 나누는 수치적 문제를 제거하는 옵션이다. 디폴트는 n=1E-12(10^{-12})이다.
NEGATIVE	비율척도, 구간척도 모형의 기울기와 로그구간 모형의 멱승이 음수가 됨을 허용한다. 로그구간 모형은 5장의 최대우도 다차원척도법을 참고 바란다.
NONORM	초기와 마지막 추정치의 정규화를 억제한다.
RANDOM	초기좌표값을 균일분포(uniformly distribution)로 부터 랜덤하게 주어지게 한다.
SINGULAR=p	0(zero)로 간주되는 값 p(0≤p≤1)를 제시한다. 디폴트는 p=1E-8이다.
UNTIE	순서척도모형 (LEVEL=ORDINAL(디폴트))에서 등위자료라도 다르게 척도하라는 옵션이다.

지금까지 PROC MDS 문에 관련된 다양한 종류의 옵션을 기능별로 살펴보았다. 독자들은 이들 옵션 중에서 대개 [부록 1.1.1]의 측정수준 지정, [부록 1.1.2]에서 자료의 형태,

[부록 1.1.3]에서 자료세트의 지정(특히 OUT= 과 OUTRES=)], [부록 1.1.4] 스트레스 및 최적척도화에 관련된 것만 잘 활용해도 PROC MDS가 제공하는 다양한 다차원척도법의 결과를 얻을 수 있을 것이다. 특히, 허명회(1994, 부록 3)의 〈PROC MDS의 용법〉은 이 절의 문법정리에 주된 역할을 하였다.

부록 1.2 ODS GRAPHICS

SAS Institute Inc.(2008)의 9.2판 이후에 ODS(Output Delivery System)를 활용한 Statistical Graphics가 PROC MDS/CORRESP/PRIN COMP/REG 등 통계적 분석에서 필요로 하는 그림을 자동으로 제공하고 있다. 이 절에서는 다차원척도법과 관련된 ODS GRAPHICS 문에 대하여 활용의 관점에서 간단히 설명하고자 한다.

먼저 ODS GRAPHICS 문은 〈ODS GRAPHICS ON;〉을 PROC MDS 문 이전에 먼저 실행시켜 다차원척도법과 관련된 그림(다차원척도그림, 스트레스-차원수 그림, 상그림, 세퍼드그림 등)을 자동으로 얻게 한다. 9.2판 이전에는 PROC MDS 문의 [부록 1.1.3] 자료세트의 지정(특히 〈OUT=과 OUTRES=〉] 옵션을 통해 얻어지는 결과를 저장한 자료를 이용하여 PROC PLOT에 의해 그림을 얻었다.

먼저 다차원척도법을 위한 ODS GRAPHICS 문의 내용을 알기 위해서는 [프로그램 1]의 템플릿(template) 절차를 실행하면 SAS의 로그 화면에 그 내용을 보여 준다. [결과 1]이 로그 화면에서 템플릿 내용만 복사한 것으로 이를 프로그램 편집 창에 가져와서 그림의 제목이나 변화를 주고 싶은 것을 편집하거나 수정하면 된다. [프로그램 2]가 원하는 형태의 다차원척도그림을 위해 수정된 PROC TEMPLATE 절차이다. 즉, [결과 1]에서 정사각형의 그림을 위해서 〈equatetype=fit〉을 〈equatetype=square〉로 수정하고, 원점을 지나는 참고 선을 넣기 위해서 〈referenceline x=0;〉과 〈referenceline y=0;〉을 추가하여 〈proc template;〉를 제일 상단에 추가한 후 SAS 프로그램 실행하듯이 하면 된다.

[프로그램 1] PROC MDS를 위한 템플릿(template) 절차

```
proc template;
        source stat.MDS.Graphics.Configuration;
```

[결과 1] PROC MDS를 위한 템플릿(template)의 내용

```
define statgraph Stat.MDS.Graphics.Configuration;
   notes "MDS Configuration Plot";
   dynamic xVar yVar Legend head;
   begingraph;
       layout overlayequated / equatetype=fit
                              xaxisopts=(offsetmin=0.1 offsetmax=0.1)
                              yaxisopts=(offsetmin=0.1 offsetmax=0.1);
         scatterplot y=YVAR x=XVAR / datalabel=LABEL;
      endlayout;
   endgraph;
end;
```

[프로그램 2] 수정된 PROC TEMPLATE 절차

```
proc template;
          define statgraph Stat.MDS.Graphics.Configuration;
   notes "MDS Configuration Plot";
   dynamic xVar yVar Legend head;
   begingraph;
       layout overlayequated / equatetype=square
                              xaxisopts=(offsetmin=0.1 offsetmax=0.1)
                              yaxisopts=(offsetmin=0.1 offsetmax=0.1);
                             referenceline x=0;
                             referenceline y=0;
         scatterplot y=YVAR x=XVAR / datalabel=LABEL;
      endlayout;

   endgraph;
end;
```

부록 1.3 PROC ALSCAL과 PROC MDS

SAS Institute Inc.(1992, pp. 272-274)는 범용 PROC ALSCAL과 PC-윈도우용 PROC MDS를 잘 비교하고 있다. 이를 간단히 요약하기로 하자. PROC MDS의 옵션으로 〈fit=2 formular=1 pfinal〉을 사용하면 PROC ALSCAL의 결과와 대동소이함을 제공하며 이는 4장의 삼원 다차원 척도법에서 활용되고 있다. 이외에는 두 절차의 결과는 때론 다를 수도 있다. PROC ALSCAL은 다차원척도법에 관련된 그림을 제공하지만 PROC MDS는 9.2판 이전에는 수행결과를 저장하여 PROC PLOT이나 PROC GPLOT을 이용하였으나 9.2판 이후에는 [부록 1.2] ODS GRAPHICS를 활용하면 된다.

〈표 3〉은 PROC ALSCAL의 옵션들 중 PROC MDS가 채택하고 있는 것을 비교하고 있다. PROC MDS의 옵션에 대해서는 이미 [부록 1.1]에서 자세히 설명하고 있다. 특히, PROC ALSCAL의 옵션 NEGATIVE (예)는 PROC MDS에 의해서 채택은 되었으나 옵션의 원래 의미와는 다름을 나타낸다.

〈표 3〉 PROC ALSCAL과 PROC MDS 옵션의 비교

PROC ALSCAL 옵션	PROC MDS 채택 여부	PROC MDS 옵션과 명령어
CONDITION=	예	
CONVERGE=	예	수렴척도는 비교가 안 됨
CUTOFF=	예	
DATA=	예	
DEGREE=	아니오	
DIMENS=	예	
DIRECTIONS=	아니오	
HEADER	예	PROC MDS에서 디폴트
IN=	예	
ITER=	예	MAXITER=
LEVEL=	예	LEVEL=NOMINAL은 지원이 안 됨
MAXDIM=m	예	DIMENSION=n TO m
MINDIM=n	예	DIMENSION=n TO m
MINSTRESS=	예	MINCRIT=
MODEL=EUCLID MODEL=INDSCAL MODEL=GEMSCAL MODEL=ASYMSCAL MODEL=ASYMINDS	예 예 아니오 아니오 아니오	COEF=IDENTITY COEF=DIAGONAL
NEGATIVE	(예)	이 옵션은 기울기나 멱승에 음수를 허용함.
NOULB	예	
OUT=	예	저장되는 내용에 차이가 있음
PLOT	아니오	
PLOTALL	아니오	
PRINT	아니오	
READV, etc.	아니오	자료세트 옵션 WHERE을 사용함
READFIXV, etc.	아니오	
ROWS=	아니오	
SHAPE=SYMMETRI SHAPE=ASYMMETR SHAPE=RECTANGU	예 예 아니오	SHAPE=TRIANGLE SHAPE=SQUARE 결측치가 있다면 SHAPE=TRIANGLE을 사용함
SIMILAR	예	
TIESTORE=	예	무시됨
UNTIE	예	

부록 1.4 PROC MLSCALE과 PROC MDS

최대우도 다차원척도법에 관한 5.1절의 끝부분에서 PROC MLSCALE과 PROC MDS의 추정방법에 있어서 차이점이 있음을 언급해 두고 있다. 여기서는 PC-윈도우용 PROC MDS가 범용 PROC MLSCALE의 어떤 옵션을 채택하고 있는지를 〈표 4〉에 간단히 요약하려 한다. 이는 SAS Institute Inc.(1992, pp. 274-275)에서 발췌한 것이다.

특별히 다음의 PROC MLSCALE과 PROC MDS는 대개 유사한 결과를 제공한다.

MDS 과정: proc mds fit=log level=loginterval … ;

MLSCALE 과정: proc mlscale stvarnce=constant suvarnce=constant… ;

〈표 4〉 PROC MLSCALE과 PROC MDS의 옵션의 관계

PROC MLSCALE 옵션	PROC MDS 채택여부	PROC MDS 옵션과 명령어
SQUARE	예	SHAPE=SQUARE
INPUT=MATRIX	아니오	디폴트
INPUT=VECTOR	아니오	
STLABEL= STLBDS	아니오 아니오	ID statement
SULABEL= SULBDS	아니오 아니오	MATRIX statement
CONFIG CONFDS=	아니오 아니오	IN=data set
NEQU= CONSDS	아니오 아니오	
METVAL METVDS	아니오 아니오	IN=
SEWGTS SEWGDS=	아니오 아니오	
SPLVAL SLPVDS=	아니오 아니오	
DIMENS=	예	
METRIC=IDENTITY METRIC=DIAGONAL METRIC=FULL	예 예 아니오	COEF=IDENTITY COEF=DIAGONAL

〈표 4(계속)〉 PROC MLSCALE과 PROC MDS의 옵션의 관계

PROC MLSCALE 옵션	PROC MDS 채택여부	PROC MDS 옵션과 명령어
TRANSFRM=SCALE TRANSFRM=POWER TRANSFRM=SPLINE	예 예 아니오	LEVEL=RATIO LEVEL=LOGINTERVAL
STVARNCE=	아니오	
SUVARNCE=	아니오	
NORMAL	아니오	디폴트(FIT=DISTANCE)
ITMAX=	예	MAXITER=
ITXMAX=	아니오	
ITWMAX=	아니오	
ITAMAX=	아니오	
ITPMAX=	아니오	
CONV=	(예)	의미가 다름(수렴의 조건을 지정)
FACTOR=	아니오	
HISTORY	아니오	PITER
ASYMP	아니오	
OUTCON	아니오	OUT=
OUTDIS	아니오	
OUTMET	아니오	OUT=
OUTSPL	아니오	
OUTRES	(예)	OUTRES=dataset

부록 2
자료

자료 1: 신체부위(파일이름: body.txt).

자료 2: 레크레이션(파일이름: recreation.txt).

자료 3: 로이드은행(파일이름: Lloyds05.sd2, Lloyds05.xlsx).

자료 4: 로이드은행(파일이름: Lloyds25.sd2, Lloyds25.xlsx).

자료 5: 켈로그(파일이름: kellogg.txt).

자료 6: 색채환(파일이름: color.txt).

자료 7: 공익회사(파일이름: utility.txt).

자료 8: 관심영역(파일이름: field.txt).

자료 9: 국가(파일이름: nation.txt).

자료 10: 모스부호(파일이름: morse.txt).

자료 11: 두개골(파일이름: skull.txt).

자료 12: 유럽 10대 도시(파일이름: eurodist.txt).

자료 13: 자동차 선호도 평가(파일이름: cartype.txt).

자료 1: 신체 부위(파일이름: body.txt).

30명(아동 15명, 성인(대학생) 15명)과 15군데 신체 부위.

cheek,	face,	mouth,	hand,	ear,	body,	arm,	elbow,	hand,	palm,	finger,	leg,	knee,	foot,	toe.
0	2	1	3	4	10	5	9	6	7	8	11	12	13	14
2	0	12	1	13	3	8	10	11	9	7	4	5	6	14
3	2	0	1	4	9	5	11	6	7	8	10	13	12	14
2	1	3	0	4	9	5	6	11	7	8	10	12	13	14
10	1	11	2	0	6	3	4	5	12	13	7	8	14	9
14	12	9	6	13	0	8	7	5	10	11	1	4	2	3
12	14	11	10	13	5	0	4	1	3	2	6	9	7	8
5	7	14	8	6	9	1	0	2	3	4	10	11	12	13
13	11	12	10	14	9	3	4	0	1	2	6	5	7	8
8	6	7	9	4	5	3	10	1	0	2	12	11	13	14
14	5	13	6	9	12	3	4	1	2	0	7	8	10	11
14	12	13	11	9	7	4	6	5	3	10	0	8	1	2
12	11	14	10	13	4	5	8	6	7	9	1	0	2	3
12	14	10	13	11	9	4	5	8	6	7	2	3	0	1
13	8	9	11	14	3	6	5	7	10	12	2	4	1	0
0	4	2	3	11	9	14	12	1	7	13	8	6	5	10
7	0	11	9	1	2	8	3	13	14	10	6	12	4	5
10	11	0	3	7	1	8	13	2	12	6	9	14	5	4
4	8	11	0	1	2	3	9	7	13	10	5	14	12	6
14	6	11	10	0	1	2	4	9	8	5	7	13	3	12
11	6	14	12	3	0	13	2	1	9	5	4	10	7	8
14	12	8	3	1	7	0	13	2	6	5	9	11	4	10
6	10	9	14	3	4	12	0	8	13	5	7	11	1	2
13	14	12	1	4	6	2	7	0	10	9	3	11	8	5
6	9	7	11	3	5	1	12	2	0	13	8	10	4	14
10	6	9	8	11	13	7	12	2	1	0	14	5	3	4
6	9	7	10	4	2	12	14	1	8	13	0	3	5	11
13	2	8	3	1	5	9	14	12	11	7	6	0	10	4
14	11	8	4	13	2	12	9	5	3	10	6	7	0	1
14	3	4	7	13	6	2	8	9	11	10	1	12	5	0
0	13	8	11	3	7	14	6	9	5	2	4	1	12	10
6	0	2	14	4	11	8	7	10	1	12	9	3	5	13
10	1	0	11	3	14	7	9	8	2	13	6	5	4	12
5	3	6	0	7	14	10	4	13	8	11	9	2	12	1
14	2	4	8	0	5	7	9	13	6	11	1	12	10	3
8	9	6	5	13	0	4	11	3	14	1	12	2	10	7
3	2	5	9	13	14	0	12	10	11	8	1	4	6	7
2	3	7	6	4	14	8	0	11	12	5	13	1	9	10
3	8	13	14	4	6	2	5	0	1	9	12	7	11	10
12	7	6	11	4	5	2	14	13	0	1	3	9	8	10
8	13	6	4	7	12	2	5	14	10	0	3	9	11	1
8	5	12	9	14	7	4	13	11	6	2	0	3	1	10
13	9	2	10	14	7	11	1	3	8	5	6	0	12	4
13	4	7	14	11	12	9	10	3	5	8	1	6	0	2
14	13	3	5	12	11	6	1	10	8	2	9	4	7	0
0	3	1	4	2	14	5	6	7	11	8	9	12	10	13
2	0	1	3	5	4	13	12	10	7	11	6	9	8	14
5	6	0	4	2	7	10	1	12	11	8	13	3	9	14
1	3	2	0	5	4	6	10	9	7	8	11	12	13	14
1	4	2	3	0	14	5	6	7	9	8	11	10	12	13
5	3	1	2	4	0	6	7	8	9	10	11	12	13	14
6	8	5	7	9	13	0	1	2	3	4	10	11	12	14
8	9	7	5	6	14	1	0	3	2	4	10	11	12	13
9	10	5	7	4	8	1	6	0	2	3	11	12	13	14
10	12	7	8	11	13	5	4	2	0	1	3	9	6	14
9	7	8	5	6	14	3	4	2	1	0	10	11	12	13
11	12	10	13	14	3	5	8	9	7	6	0	4	1	2
7	9	6	5	8	4	10	12	13	11	14	1	0	3	2
6	4	7	5	13	14	9	8	10	11	12	1	3	0	2
11	8	10	9	14	4	5	12	6	13	7	1	3	2	0
0	12	10	1	11	13	7	4	2	9	14	8	3	6	5
3	0	1	2	4	12	8	11	9	13	5	7	14	6	10
4	3	0	1	11	8	5	10	9	6	7	13	12	2	14
9	2	11	0	1	8	14	3	4	10	6	7	13	12	5
12	13	5	1	0	14	6	11	9	3	4	2	7	10	8
9	10	5	13	7	0	12	6	2	11	1	3	4	14	8
7	8	10	12	13	14	0	4	1	2	3	5	9	11	6

9	4	14	5	2	7	11	0	10	12	3	6	1	8	13
7	9	14	13	6	11	2	3	0	1	12	10	4	8	5
9	6	10	3	8	12	4	13	5	0	11	14	7	1	2
6	9	11	10	4	13	5	2	1	3	0	7	8	12	14
7	2	3	13	14	6	9	10	4	11	12	0	5	1	8
11	14	4	2	13	8	7	1	9	12	10	6	0	5	3
8	11	14	10	1	7	12	9	4	13	6	3	5	0	2
7	10	9	14	11	12	5	4	6	13	8	2	3	1	0
0	1	4	2	3	9	10	11	12	13	14	5	6	7	8
3	0	2	1	4	5	9	8	13	10	11	14	12	6	7
3	1	0	2	9	11	4	5	7	6	8	10	12	13	14
14	1	2	0	4	8	5	7	3	13	6	10	9	11	12
4	1	3	2	0	12	6	9	5	7	8	10	11	13	14
13	2	3	1	14	0	4	5	12	6	7	8	9	10	11
14	11	12	13	9	10	0	1	8	2	3	7	4	5	6
14	10	11	12	13	9	1	0	2	3	4	6	5	7	8
13	10	11	12	14	8	3	7	0	1	2	5	6	4	9
12	10	11	9	14	13	2	4	3	0	1	6	7	5	8
14	11	10	12	9	13	4	5	3	2	0	6	8	7	1
11	10	12	9	13	14	4	5	6	7	8	0	1	2	3
10	13	11	9	14	8	7	4	12	5	6	1	0	2	3
13	9	11	10	12	14	5	8	4	6	7	2	3	0	1
12	13	11	9	14	4	10	8	6	7	5	2	3	1	0
0	11	13	14	1	2	5	10	3	6	4	8	7	9	12
2	0	12	9	13	4	14	8	11	6	1	7	10	5	3
11	3	0	8	13	6	12	2	9	7	5	10	1	14	4
13	11	14	0	7	9	6	1	10	12	8	4	3	5	2
10	11	13	7	0	4	14	2	9	1	12	3	8	5	6
4	9	11	7	3	0	14	5	10	2	6	12	13	1	8
3	9	11	7	10	8	0	14	5	6	13	2	12	1	4
11	14	9	13	3	10	5	0	7	2	1	4	6	8	12
5	13	2	9	8	7	4	14	0	3	12	10	11	6	1
10	7	3	5	1	6	12	8	11	0	14	13	9	2	4
9	1	11	12	4	10	5	2	14	6	0	8	13	7	3
10	1	7	11	4	3	9	12	5	8	2	0	14	6	13
4	3	11	7	5	13	2	12	8	1	14	6	0	10	9
1	9	6	2	7	11	14	3	10	12	8	5	4	0	13
12	6	14	1	8	9	13	7	4	3	11	5	10	2	0
0	14	13	11	9	7	5	10	3	1	2	8	12	4	6
14	0	11	13	7	3	2	6	4	12	5	9	10	1	8
5	8	0	13	2	9	4	7	12	3	10	1	14	11	6
4	13	11	0	12	8	10	9	5	1	3	7	2	14	6
11	12	6	9	0	7	5	8	13	10	4	2	14	1	3
5	14	3	7	12	0	13	6	8	2	9	10	11	4	1
2	11	8	7	4	14	0	13	12	10	5	1	3	9	6
2	11	13	14	4	8	10	0	7	1	9	3	5	12	6
8	12	7	6	9	10	11	14	0	1	2	4	5	13	3
7	11	10	14	13	3	12	4	6	0	5	2	8	9	1
8	12	13	9	14	4	3	11	1	2	0	5	6	7	10
4	9	11	10	5	3	6	7	13	8	14	0	12	2	1
9	12	10	6	3	14	5	11	13	7	8	4	0	2	1
6	4	12	14	7	3	11	10	5	8	1	2	13	0	9
4	2	5	11	13	9	1	10	12	3	6	7	14	8	0
0	1	3	2	4	10	5	6	7	8	9	11	12	13	14
3	0	2	1	4	5	10	11	13	12	14	6	7	8	9
1	2	0	3	4	10	5	6	8	7	9	11	12	13	14
4	1	3	0	2	5	6	7	8	9	11	10	12	13	14
3	2	4	1	0	5	7	6	9	8	10	11	12	13	14
12	11	13	10	14	0	5	6	8	7	9	1	2	3	4
13	11	12	10	14	5	0	1	3	2	4	7	6	8	9
8	6	7	5	9	10	1	0	3	2	4	12	11	13	14
13	11	14	10	12	9	3	4	0	1	2	6	7	5	8
7	5	6	8	9	10	3	4	1	0	2	11	12	13	14
7	6	8	5	9	10	3	4	1	2	0	11	12	13	14
13	5	6	4	14	3	12	7	8	9	10	0	11	1	2
13	10	12	9	14	4	5	6	7	11	8	1	0	2	3
8	6	9	5	7	4	10	11	13	12	14	2	3	0	1

12	13	11	5	14	4	6	7	9	8	10	2	3	1	0
0	7	2	12	4	14	5	3	13	9	10	6	1	11	8
7	0	5	3	13	9	11	4	10	12	8	1	14	6	2
12	13	0	10	6	3	1	14	9	11	4	5	8	7	2
1	11	5	0	7	8	13	10	6	2	9	14	12	3	4
7	8	1	10	0	6	14	13	12	3	9	5	11	2	4
13	14	2	10	11	0	3	12	1	9	8	5	7	6	4
1	14	6	7	11	10	0	12	13	4	2	3	5	8	9
14	5	2	9	13	1	6	0	8	10	12	4	11	7	3
4	9	5	7	13	8	14	6	0	3	2	10	1	11	12
9	13	3	1	10	12	8	7	6	0	11	4	2	14	5
3	7	9	5	14	8	11	4	12	6	0	2	13	1	10
5	6	13	14	11	9	8	4	10	7	2	0	1	3	12
9	14	10	12	11	13	5	2	8	7	6	1	0	3	4
14	7	13	2	11	12	4	6	8	3	10	1	9	0	5
5	1	8	3	11	4	9	6	13	12	14	10	7	2	0
0	2	1	4	3	5	6	7	9	8	10	11	12	13	14
2	0	1	4	3	10	5	6	8	7	9	11	12	13	14
1	4	0	3	2	5	9	10	7	6	8	11	12	13	14
3	1	2	0	4	10	5	6	8	7	9	11	12	13	14
1	3	2	4	0	7	5	6	8	9	10	11	12	13	14
4	2	3	1	5	0	6	7	8	9	10	11	12	13	14
8	6	7	5	9	10	0	1	4	2	3	11	12	13	14
8	6	7	5	9	10	1	0	2	3	4	11	12	13	14
6	5	8	9	7	10	3	4	0	1	2	11	12	13	14
8	5	7	6	9	10	4	3	2	0	1	11	12	13	14
6	9	5	8	7	10	3	4	1	2	0	11	12	13	14
6	8	5	9	7	4	10	11	13	12	14	0	1	2	3
6	4	5	3	7	2	8	9	10	11	12	1	0	13	14
8	6	7	5	9	4	10	11	13	12	14	2	3	0	1
8	5	6	9	7	4	10	11	12	13	14	3	2	1	0
0	7	10	8	12	6	13	1	9	2	14	5	3	11	4
2	0	13	1	11	6	14	5	9	3	12	8	4	10	7
10	9	0	12	5	8	2	11	13	14	6	3	4	1	7
9	1	14	0	7	13	6	4	2	5	10	11	3	12	8
5	2	14	1	0	9	13	10	4	8	11	6	7	3	12
13	1	14	2	4	0	12	5	8	3	9	10	11	6	7
14	12	8	10	7	3	0	11	6	13	4	1	9	2	5
1	6	13	5	4	14	7	0	10	2	9	12	3	11	8
6	7	14	8	3	11	13	5	0	1	10	12	4	2	9
1	10	14	9	12	13	7	3	6	0	8	11	2	5	4
8	10	13	9	6	14	3	7	5	12	0	4	11	2	1
10	13	14	9	11	12	1	7	5	8	4	0	6	2	3
11	13	14	12	10	9	8	2	7	3	5	1	0	6	4
14	5	7	4	11	3	13	10	1	8	2	12	6	0	9
2	9	14	11	12	13	10	3	8	5	1	7	6	4	0
0	3	1	4	2	5	6	7	8	9	10	11	12	13	14
2	0	3	1	4	5	6	7	8	9	10	11	12	13	14
1	3	0	2	4	5	6	7	8	9	10	11	12	13	14
5	1	2	0	3	4	6	7	8	9	10	11	12	13	14
1	10	2	11	0	3	4	5	6	7	8	9	12	13	14
13	11	12	10	14	0	1	2	3	4	5	6	7	8	9
13	11	12	10	14	7	0	1	2	3	4	5	6	8	9
12	9	11	10	14	3	5	0	4	6	7	2	1	8	13
13	10	12	11	14	5	3	4	0	1	2	8	9	6	7
13	10	12	11	14	5	3	4	2	0	1	6	7	8	9
13	11	12	10	14	9	5	3	7	4	0	6	8	2	1
13	10	11	9	12	8	1	14	6	7	5	0	2	3	4
10	12	9	14	13	4	5	11	6	7	8	1	0	2	3
13	11	12	10	14	3	4	9	1	2	8	5	7	0	6
13	11	12	7	14	4	9	10	8	6	5	2	3	1	0
0	6	1	5	2	7	8	4	9	3	10	13	11	14	12
9	0	8	1	10	5	13	3	2	11	14	6	4	7	12
6	1	0	9	5	14	4	3	12	2	7	10	13	8	11
7	1	3	0	14	9	10	11	6	4	13	2	5	12	8
13	5	4	8	0	3	11	9	7	1	2	10	12	14	6
13	8	11	2	10	0	14	12	6	5	7	1	4	3	9
12	3	14	9	13	4	0	11	10	6	5	1	8	2	7
13	14	6	12	2	7	3	0	10	11	4	9	1	8	5
7	6	13	11	12	10	3	8	0	2	1	14	9	5	4
2	11	12	9	3	4	5	13	1	0	14	8	6	10	7
13	10	6	14	7	4	9	12	2	5	0	8	11	3	1
11	13	7	12	8	4	1	6	9	14	3	0	10	5	2
3	9	4	12	6	13	8	5	11	2	10	1	0	7	14
5	10	9	11	8	14	7	3	4	12	6	1	13	0	2
3	11	5	8	9	14	12	4	10	2	13	6	7	1	0
0	1	3	2	4	5	6	7	9	14	10	11	8	12	13
3	0	4	1	2	6	13	9	14	11	10	5	7	8	12
2	3	0	8	5	1	12	4	11	13	10	6	14	9	7
4	3	2	0	1	14	7	5	6	9	8	10	11	12	13
9	1	7	6	0	5	2	3	8	10	13	11	12	14	4
1	6	3	10	7	0	5	2	12	8	9	13	4	11	14
8	9	12	3	6	11	0	4	1	14	2	7	13	10	5
9	5	4	6	7	2	1	0	3	14	11	10	13	12	8
14	8	9	7	13	3	11	4	0	5	1	6	12	10	2
11	6	14	4	7	3	13	12	1	0	2	10	5	9	8
12	10	8	5	9	11	4	2	1	3	0	7	14	6	13
6	10	4	8	14	3	7	9	13	12	11	0	5	1	2

7	5	12	14	2	13	9	11	4	8	6	1	0	3	10
7	6	5	10	9	8	13	14	11	2	12	4	3	0	1
14	11	8	7	6	4	12	3	10	13	9	2	5	1	0
0	1	2	3	4	9	7	10	6	5	8	12	11	13	14
2	0	1	3	4	8	6	10	12	5	11	7	9	13	14
2	1	0	3	4	6	8	12	7	10	5	9	11	13	14
3	1	2	0	4	5	6	7	10	8	9	12	11	13	14
3	2	4	1	0	5	9	10	7	8	6	11	12	13	14
3	2	6	1	8	0	7	10	12	9	14	4	11	5	13
13	11	12	10	14	4	0	1	2	9	3	5	7	6	8
11	12	13	10	14	5	1	0	2	9	3	6	4	7	8
13	8	10	11	14	9	3	5	0	1	2	12	7	4	6
12	11	7	13	14	10	6	4	1	0	2	8	9	3	5
10	11	6	8	7	5	9	14	1	2	0	12	13	4	3
13	9	12	10	14	4	5	11	6	7	8	0	1	2	3
13	12	11	10	14	9	6	2	5	8	7	1	0	3	4
13	11	14	10	12	6	7	9	4	8	5	2	3	0	1
13	12	14	10	11	5	8	9	6	7	3	2	4	1	0
0	4	2	9	3	14	12	6	10	1	5	13	8	11	7
3	0	8	9	6	14	10	12	1	7	4	11	13	2	5
2	3	0	4	1	14	11	10	9	7	6	13	12	8	5
13	12	11	0	10	14	4	6	1	8	5	2	7	3	9
3	4	1	5	0	14	12	11	10	6	2	13	8	9	7
11	3	6	5	13	0	1	9	7	12	4	2	10	8	14
12	13	7	2	10	14	0	5	3	11	8	1	5	4	9
13	10	8	11	9	14	6	0	4	12	2	7	1	5	3
5	11	7	13	12	14	2	10	0	4	6	3	9	1	8
1	5	2	11	3	14	12	8	6	0	4	13	9	10	7
11	12	9	13	10	14	6	4	2	5	0	8	7	3	1
14	11	12	2	13	5	1	9	4	8	10	0	6	3	7
10	12	9	13	11	14	7	1	5	8	3	6	0	4	2
12	9	10	13	11	14	8	7	1	4	3	6	5	0	2
11	12	9	13	10	14	8	3	7	4	1	6	2	5	0
0	2	1	4	3	5	9	10	6	7	8	14	11	13	12
3	0	2	1	4	5	7	14	6	10	8	9	12	11	13
1	2	0	3	4	9	13	11	8	10	5	6	14	12	7
4	1	2	0	3	5	7	10	8	14	12	6	9	11	13
2	3	4	1	0	6	8	12	5	9	7	11	13	14	10
13	4	6	1	10	0	3	9	8	14	11	2	5	7	12
13	12	10	14	11	5	0	2	4	3	8	1	6	7	9
11	13	8	10	14	9	4	0	2	12	5	7	1	6	3
14	10	8	13	9	7	5	6	0	2	1	12	11	3	4
5	10	11	14	13	8	3	4	1	0	2	12	9	6	7
13	12	10	14	9	11	6	2	4	7	0	5	3	8	1
14	13	11	9	12	10	4	5	6	7	8	0	1	2	3
13	14	9	12	10	4	8	1	7	11	5	2	0	3	6
14	13	11	10	9	6	8	12	4	7	5	2	3	0	1
14	12	6	13	10	7	5	11	8	9	4	2	3	1	0
0	2	3	9	5	14	6	13	4	1	10	7	12	8	11
9	0	1	2	12	11	3	14	6	10	5	4	13	8	7
10	1	0	8	9	12	4	13	2	11	3	5	14	7	6
13	3	6	0	7	14	1	10	4	12	8	2	11	5	9
6	3	1	4	0	14	9	11	8	7	2	10	12	5	13
7	1	2	9	10	0	11	12	3	8	4	5	13	6	14
13	7	8	2	9	14	0	10	3	12	5	1	11	4	6
13	5	4	8	11	14	6	0	3	12	1	7	2	10	9
9	4	6	5	11	14	2	13	0	10	7	3	12	1	8
1	3	2	6	10	4	7	12	5	0	11	8	13	9	14
13	9	4	10	11	14	7	2	5	12	0	8	3	6	1
6	1	5	2	10	9	3	14	4	7	11	0	13	8	12
12	11	9	13	10	14	6	1	7	8	3	5	0	4	2
11	12	6	13	9	14	3	8	1	10	5	2	7	0	4
13	7	2	10	11	14	6	9	4	12	1	5	8	3	0
0	1	2	3	4	6	7	14	8	5	9	11	10	13	12
2	0	3	1	4	6	10	9	7	5	13	11	8	12	14
5	1	0	2	4	3	7	11	8	14	9	6	10	13	12
5	2	4	0	3	1	7	10	9	14	12	6	11	8	13
12	2	3	1	0	4	5	13	7	14	9	6	10	8	11
13	8	6	1	7	0	3	10	5	14	11	2	9	4	12
14	13	10	12	11	6	0	2	3	7	4	1	5	8	9
14	13	9	10	12	6	2	0	3	11	4	8	1	7	5
14	13	9	11	12	7	3	6	0	4	2	10	8	1	5
4	9	7	12	8	5	11	10	1	0	3	14	13	2	5
13	14	10	12	11	9	7	4	2	3	0	8	5	6	1
13	12	10	11	9	3	1	6	7	14	8	0	5	2	4
12	13	9	11	10	7	8	1	6	14	3	2	0	5	4
13	14	10	12	11	7	9	6	1	8	5	3	4	0	2
14	12	8	11	10	9	7	2	5	13	1	6	3	4	0
0	1	2	3	5	4	10	13	6	7	8	11	14	12	9
2	0	4	1	5	3	7	12	6	8	10	9	13	11	14
3	1	0	2	9	7	8	13	4	10	5	11	14	6	12
4	1	2	0	6	3	9	13	5	8	7	12	14	11	10
5	4	6	1	0	3	13	11	8	9	2	14	12	10	7
10	4	6	1	13	0	2	12	5	7	8	3	11	9	14
13	12	11	10	14	3	0	6	2	4	5	1	9	7	8
13	11	12	10	14	3	2	0	4	6	5	8	1	7	9
13	9	8	7	14	6	4	5	0	2	3	10	11	1	12
8	5	6	10	14	3	4	9	1	0	2	12	13	7	11

12	13	4	14	5	6	7	8	1	3	0	11	10	9	2
13	11	12	10	14	5	1	7	6	9	8	0	2	3	4
11	13	12	14	9	3	6	1	7	10	8	2	0	4	5
13	12	7	5	14	4	9	8	1	11	10	3	6	0	2
12	13	9	14	11	4	8	10	5	7	1	3	6	2	0
0	1	3	2	4	5	9	10	7	6	11	12	13	8	14
2	0	1	3	4	5	11	13	7	6	9	12	14	8	10
1	2	0	3	4	5	11	12	8	7	6	13	14	10	9
2	1	4	0	3	5	6	13	7	10	12	8	14	9	11
1	3	2	4	0	14	8	12	7	5	6	11	13	10	9
11	6	7	1	8	0	2	14	4	12	9	3	13	6	10
13	12	14	11	9	10	0	6	3	8	2	1	7	4	3
13	12	9	11	14	10	2	0	3	4	5	6	1	7	8
12	13	9	14	10	11	5	7	0	3	2	6	8	1	4
9	11	13	12	10	14	5	7	1	0	2	6	8	4	3
12	11	13	10	9	14	4	5	3	2	0	8	7	6	1
12	14	13	10	11	9	1	4	5	8	7	0	2	3	6
14	12	9	11	13	10	4	1	5	8	7	2	0	3	6
13	14	10	12	9	11	6	8	1	4	3	5	7	0	2
13	14	11	9	10	12	6	8	3	4	1	5	7	2	0
0	1	2	4	3	5	6	7	8	9	10	11	12	13	14
1	0	3	2	4	5	10	11	12	13	14	6	7	8	9
2	1	0	4	3	5	6	7	8	9	10	11	12	13	14
2	1	3	0	4	5	10	11	12	13	14	6	7	8	9
3	2	4	1	0	5	6	7	8	9	10	11	12	13	14
3	2	4	1	5	0	6	7	8	9	10	11	12	13	14
12	10	11	9	13	14	0	1	2	3	4	5	6	7	8
10	9	11	8	12	13	1	0	3	6	7	5	2	4	14
11	10	12	14	13	9	3	4	0	1	2	7	8	5	6
11	10	12	9	14	13	3	4	1	0	2	7	8	5	6
10	12	11	14	13	9	4	3	2	1	0	8	7	6	5
12	11	13	10	14	9	4	5	6	7	8	0	1	2	3
11	12	10	14	13	4	6	5	7	8	9	1	0	2	3
12	10	11	14	13	9	8	7	4	5	6	3	2	0	1
11	13	10	14	12	4	9	8	7	6	5	2	3	1	0
0	1	2	4	3	5	6	7	8	9	10	11	12	13	14
2	0	1	4	3	5	6	7	8	9	10	11	12	13	14
1	3	0	4	2	14	9	8	7	6	5	13	12	11	10
3	1	2	0	4	14	5	6	7	8	9	10	11	12	13
1	3	2	4	0	5	6	11	8	10	14	7	12	9	13
4	2	3	1	5	0	6	7	8	9	10	11	12	13	14
12	10	11	14	13	5	0	1	2	3	4	6	7	8	9
9	11	10	13	12	14	2	0	3	5	4	8	1	7	6
12	10	11	9	13	14	4	3	0	2	1	8	7	5	6
13	11	12	10	14	9	5	6	1	0	3	7	8	2	4
10	12	9	13	11	14	5	2	4	8	0	7	3	6	1
13	11	12	10	14	9	4	5	6	7	8	0	1	2	3
13	11	12	10	14	9	5	4	6	7	8	1	0	2	3
12	9	10	13	11	14	7	5	1	8	3	6	4	0	2
10	12	9	13	11	14	7	2	6	8	1	5	3	4	0
0	2	3	4	5	1	6	7	8	9	10	11	12	13	14
3	0	4	2	5	1	6	7	8	9	10	11	12	13	14
3	2	0	5	4	1	6	7	8	9	10	11	12	13	14
4	3	5	0	2	1	6	7	8	9	10	11	12	13	14
4	3	5	2	0	1	10	9	6	7	8	14	13	12	11
2	4	3	1	5	0	6	7	8	9	10	11	12	13	14
13	12	14	10	11	1	0	2	3	4	5	6	7	8	9
13	12	14	10	11	1	2	0	3	4	5	7	6	8	9
11	10	12	14	13	1	4	5	0	2	3	8	9	6	7
12	10	11	14	13	1	4	5	2	0	3	6	7	8	9
11	12	10	14	13	1	4	5	2	3	0	8	9	7	6
13	12	14	10	11	1	5	6	7	8	9	0	2	3	4
11	10	12	14	13	1	6	5	7	8	9	2	0	3	4
11	10	12	14	13	1	9	5	8	7	6	2	4	0	3
12	11	13	14	10	1	8	9	6	7	5	3	4	2	0
0	3	1	4	2	14	9	8	7	6	5	13	12	11	10
2	0	1	4	3	14	5	6	9	7	8	13	12	11	10
1	2	0	4	3	14	5	6	9	7	8	13	12	11	10
3	1	4	0	2	14	5	6	7	8	9	10	11	12	13
1	3	2	4	0	14	9	8	7	6	5	11	10	13	12
9	2	10	1	8	0	4	11	3	5	12	6	13	7	14
6	9	7	5	8	14	0	1	2	3	4	10	11	12	13
5	8	6	9	7	14	1	0	4	3	2	11	10	13	12
12	9	11	13	10	14	4	3	0	1	2	5	8	7	6
10	12	11	13	9	14	4	3	2	0	1	8	7	6	5
9	11	10	13	12	14	4	3	2	1	0	8	7	6	5
9	11	10	13	12	14	8	7	6	5	4	0	1	3	2
9	11	10	13	12	14	8	7	6	5	4	1	0	3	2
12	10	11	9	13	14	4	5	6	7	8	2	3	0	1
12	10	11	9	13	14	7	8	4	5	6	2	3	1	0
0	1	5	7	6	14	9	4	8	2	11	10	3	12	13
2	0	10	5	9	8	6	14	3	1	11	7	13	4	12
2	5	0	4	1	14	12	8	6	10	3	11	9	13	7
9	8	12	0	11	1	2	13	5	10	6	3	14	4	7
7	10	3	13	0	14	8	12	4	6	2	9	11	5	1
7	2	12	1	11	0	3	13	6	8	9	4	14	5	10
7	5	13	2	10	14	0	11	3	6	8	1	12	4	9

5	12	2	13	3	14	10	0	8	4	6	11	1	9	7
7	2	13	5	14	12	3	8	0	6	10	4	9	1	11
1	4	14	7	11	8	5	12	2	0	9	6	13	3	10
7	5	12	13	2	14	8	10	3	6	0	9	11	4	1
7	2	13	3	14	4	1	12	6	8	10	0	11	5	9
7	12	2	13	3	14	11	1	8	6	5	10	0	9	4
2	8	11	13	5	14	7	10	1	12	4	6	9	0	3
11	12	10	13	2	14	6	8	4	9	1	5	7	3	0
0	1	2	3	4	14	8	10	5	6	7	13	11	9	12
2	0	3	1	4	5	8	10	6	7	12	13	9	11	14
1	2	0	4	3	14	11	13	7	5	6	9	10	12	8
3	1	2	0	4	14	8	10	6	5	7	9	11	13	12
2	4	1	3	0	14	5	12	6	8	7	10	13	11	9
13	9	11	10	14	0	2	4	5	12	8	1	3	6	7
10	12	13	11	9	14	0	7	2	4	3	1	6	5	8
10	13	9	12	11	14	2	0	3	4	5	6	1	7	8
12	13	10	9	11	14	4	5	0	1	2	6	8	3	7
11	10	12	9	13	14	3	4	1	0	2	8	6	5	7
10	12	9	13	8	14	4	11	2	3	0	6	7	5	1
8	7	12	13	10	14	3	6	5	11	9	0	4	1	2
11	12	9	13	10	14	5	1	6	8	7	2	0	3	4
11	12	9	13	10	14	6	7	1	5	8	3	4	0	2
11	10	12	13	9	14	6	7	5	8	1	3	4	2	0
0	1	3	4	2	14	12	10	6	5	8	13	11	7	9
3	0	2	1	4	14	8	10	6	5	12	9	11	7	13
2	1	0	3	4	14	9	5	7	12	11	10	6	8	13
3	1	2	0	4	14	5	9	7	11	12	6	10	8	13
1	4	2	3	0	14	12	10	8	7	5	13	11	9	6
13	2	14	1	12	0	4	6	7	9	10	3	5	8	11
12	10	11	9	13	14	0	2	4	5	7	1	3	6	8
12	10	9	13	11	14	2	0	3	7	6	4	1	5	8
12	9	11	10	13	14	5	4	0	1	2	8	7	3	6
4	12	7	13	6	14	9	8	1	0	2	11	10	3	5
7	12	6	13	5	14	9	8	3	2	0	11	10	4	1
11	10	12	9	13	14	4	5	6	7	8	0	1	2	3
12	10	9	13	11	14	3	1	5	8	7	2	0	4	6
11	10	12	9	13	14	7	6	1	4	8	3	5	0	2
10	12	11	13	9	14	8	7	4	5	1	3	6	2	0
0	1	2	4	3	14	12	10	6	5	8	13	11	7	9
2	0	3	1	4	14	8	10	5	6	12	9	11	7	13
2	1	0	4	3	14	12	9	5	11	7	13	10	6	8
2	1	4	0	3	14	8	10	6	11	12	7	9	5	13
1	3	2	4	0	14	12	9	8	7	5	13	10	11	6
13	10	11	1	14	0	3	7	5	12	8	2	6	4	9
13	9	11	10	12	14	0	1	3	6	5	2	4	7	8
10	12	8	13	9	14	1	0	6	11	4	3	2	5	7
12	9	11	10	13	14	5	4	0	1	3	8	6	2	7
9	12	11	13	10	14	5	4	1	0	2	6	7	3	8
10	12	11	13	9	14	6	5	1	3	0	8	7	4	2
12	10	11	9	13	14	2	5	6	8	7	0	1	3	4
11	12	9	13	10	14	3	2	7	8	4	1	0	5	6
12	10	11	9	13	14	5	8	1	6	7	3	4	0	2
10	12	11	13	9	14	8	7	4	6	2	3	5	1	0

자료 2: 레크레이션(파일이름: recreation.txt).

10명의 15가지 레크리에이션 평가.

concert, museum, theatre, movie, watch TV, conference, reading, watch hockey, ballet, political debate, fashion show, documentary film, exhibition, window shopping, restaurant.

```
 0
 8  0
18 21  0
 5 22 16  0
 8 24 22 11  0
11 13 19 18 22  0
19 12 15  9 21 14  0
 6  6 15 19 21 13 23  0
24 18 21 15  9 19 16 23  0
21 21 23 23 23 13 12 11 23  0
22 20 24 21 23 10 13 12  7  7  0
10  7 21 11  9 10  6  8 14 22 22  0
23 21 18  9  8 10  7 14 22  9 11  8  0
 5 14 19 16 23 22 19 20  8 17 23 21 20  0
14 23 11  8 21 22  7 10 23 19 21  9  7 22  0
 0
 9  0
23 21  0
16 17 12  0
14 19 20 19  0
 7 21 18 15 19  0
11 12 11 17 22  6  0
 6  5 15 14 17  3 18  0
21 15 22 12 12 23 10 21  0
23 24 21 20 17 18 18 15 20  0
18 14 17 18 18 10 12 14  7  9  0
 2 10 16 16  6  5 17 15 22 16 10  0
18  4  5  4 12 16  8 16 20 22 19  9  0
16  9 18  6 11 17  4  8  7 12 14 13 18  0
17 18  2  2 17 14 10 16 20  7 17 11  5 22  0
```

```
0
7  0
21 12  0
17 23 11  0
 9 24 12 25  0
 5 22 18 10 24  0
 8 10 13 21 23  7  0
10  3 17 23 24 18 25  0
25 15 20  7 12 20 11 22  0
18 24 22 24 19  7  5 13 25  0
25 19 25 25 25  4 11  8  4  8  0
 3  4  8 10  7 11 16 20 22 21 19  0
23 22  5  7 10  2 11 11 10 12 14 11  0
10  5 18 21 17 24  9  7 11 13 19  8 22  0
10 11  4  6 11 15  3  7 11 10 20 10  7  9  0
0
10  0
16 20  0
 5 11  9  0
18 21 17 19  0
 3  7 19 22 18  0
13  1 16 20 16  7  0
 9 15 20 17 19 13 22  0
24 18 21  5 10 21  7 18  0
22 17 23 18 16 21 24 25 21  0
18 21 17 15 14 21 18 16  3 22  0
 3 18 21 23 15 12 23  3 20 13 21  0
23 11 18 15 13 21 24 18 22 13 20 11  0
13 19 23 24 22 15  4 15 10 21  7  5 15  0
21 25 12  9 22 13  6 17 22  7 19 18 24 25  0
```

0
9 0
22 20 0
5 20 18 0
9 22 18 16 0
12 18 24 22 18 0
22 8 16 20 23 19 0
3 3 10 12 14 4 16 0
22 24 22 3 9 23 15 10 0
22 22 22 24 17 15 15 23 9 0
6 23 23 23 12 21 21 15 4 12 0
4 19 15 20 12 17 4 11 7 20 10 0
22 7 12 16 19 11 19 13 22 19 19 9 0
12 10 23 23 7 15 11 11 20 17 15 17 14 0
21 7 4 4 4 9 4 17 25 8 19 16 16 25 0
0
15 0
10 23 0
5 20 16 0
16 22 20 15 0
7 9 8 15 19 0
17 8 17 6 16 22 0
3 4 10 21 23 10 20 0
18 20 11 7 12 21 8 7 0
22 23 19 18 13 16 16 20 22 0
6 18 20 21 21 5 16 15 7 16 0
22 13 9 18 9 16 7 7 15 17 18 0
21 6 17 7 9 18 10 16 9 11 18 7 0
8 17 21 17 8 8 17 18 15 22 20 20 19 0
20 20 7 6 16 17 8 14 22 10 18 14 17 24 0
0
12 0
22 14 0
7 19 7 0
18 22 22 15 0
11 16 12 16 23 0
16 12 15 22 22 13 0
7 6 13 21 25 13 20 0
25 19 13 9 19 23 19 19 0
25 12 21 21 15 10 8 12 19 0
22 20 20 22 25 12 13 12 3 8 0
8 8 21 9 10 9 9 13 21 10 22 0
25 22 9 18 18 22 7 12 19 7 10 16 0
13 21 18 23 23 23 7 23 13 21 15 8 22 0
15 25 2 7 18 6 3 12 25 12 21 16 5 22 0

0
8 0
12 16 0
14 19 12 0
8 21 18 16 0
6 14 16 7 18 0
15 8 17 13 19 17 0
9 11 18 19 16 9 20 0
21 15 17 6 15 18 14 16 0
20 21 22 23 20 21 21 21 17 0
9 17 18 20 21 13 11 11 10 24 0
7 23 14 22 8 9 22 8 22 19 23 0
24 12 11 18 8 7 21 17 14 9 13 8 0
7 6 23 22 5 16 8 7 9 10 12 15 19 0
19 6 4 5 7 16 2 19 17 13 12 11 25 23 0
0
7 0
7 18 0
4 10 4 0
4 20 22 15 0
2 5 9 3 16 0
9 2 12 20 15 12 0
10 6 8 18 10 5 7 0
20 21 12 5 8 19 3 10 0
8 22 19 22 14 11 6 6 14 0
8 15 16 17 19 10 13 7 9 10 0
10 8 16 10 6 6 16 4 17 10 19 0
22 22 8 6 6 6 8 12 8 10 11 8 0
10 9 20 21 7 19 9 11 18 22 13 10 19 0
22 13 2 3 16 7 2 20 22 12 22 15 20 23 0
0
22 0
25 9 0
3 22 25 0
14 25 16 7 0
18 22 7 20 24 0
22 5 11 14 22 5 0
9 6 11 22 22 4 19 0
25 21 18 4 13 18 14 22 0
22 21 23 16 19 20 21 20 12 0
19 22 23 21 23 7 21 15 7 18 0
4 19 23 21 12 22 7 6 19 21 12 0
9 6 16 12 18 22 22 13 21 20 21 6 0
21 6 22 22 5 23 4 22 17 19 23 20 17 0
16 22 17 4 7 5 16 16 21 6 21 5 23 17 0

자료 3: 로이드은행(파일이름: Lloyds05.sd2, Lloyds05.xlsx).

1905년~1909년 근무한 80명 은행원 : 515-4514.

515	0							
547	4.214285714	0						
573	3.333333333	4.904761905	0					
638	4.108108108	0.892857143	4.837837838	0				
653	4.634146341	2.178571429	4.87804878	2.682926829	0			
781	3.357142857	4.797619048	1.791666667	4.567567568	5.182926829	0		
794	3.511627907	3.406976744	4.953488372	3.697674419	3.279069767	4.965116279	0	
797	4.243902439	0.744047619	4.597560976	1.097560976	1.682926829	4.756097561	·····	
842	3.404761905	4.952380952	1.875	4.878378378	5.158536585	1.045454545	·····	
883	3.795454545	1.886363636	5.102272727	2	3.795454545	4.875	·····	
914	4.31875	1.291666667	4.5875	1.5	1.487804878	4.9875	·····	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
4514	3.464285714	1.238095238	4.836956522	1.488095238	3.059523809	4.797619047	·····	0

자료 4: 로이드은행(파일이름: Lloyds25.sd2, Lloyds25.xlsx).

1925년~1929년 근무한 80명 은행원 : 505-4570.

505	0							
509	0.866666667	0						
518	3.922222222	3.651162791	0					
630	6.233333333	6.197674419	5.85	0				
631	4.966666667	4.825581395	2.95	5.416666667	0			
735	1.733333333	1.61627907	3.75	6.214285714	4.857142857	0		
826	1.094444444	1.405555556	3.933333333	6.211111111	5.027777778	1.7		
828	1.133333333	0.412790698	3.648809524	6.202380952	4.803571429	1.416666667	·····	
886	6.066666667	6.046511628	5.525	2.5	4.875	6.035714286	·····	
893	2.144444444	2.465116279	3.61627907	6.197674419	4.790697674	2.11627907	·····	
1031	2.855555556	2.697674419	3.357142857	6.185714286	4.7	1.928571429	·····	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
4570	6.277777778	6.220930233	5.9	1	5.583333333	6.19047619	·····	0

자료 5: 켈로그(파일이름: kellogg.txt).

23종류의 시리얼과 10개의 측정변수.

X_1: 칼로리, X_2: 단백질(g), X_3: 지방(g), X_4: 나트륨(mg), X_5: 다이어트 식이섬유(g), X_6: 복합 탄수화물(g), X_7: 당분(g), X_8: 칼륨(mg), X_9: 비타민과 무기물(하루 권장량(%) : 0, 25, 100), X_{10}: 유형(온 또는 냉).

AllB	0.1818	0.6	0.3333	0.8125	0.6429	0	0.3333	1	0.9677	0
AllF	0	0.6	0	0.4375	1	0.0667	0	1	1	0
AppJ	0.5455	0.2	0	0.3906	0.0714	0.2667	0.9333	0.5	0.0323	0
CorF	0.4545	0.2	0	0.9063	0.0714	0.9333	0.1333	0	0.0484	0
CorP	0.5455	0	0	0.2813	0.0714	0.4	0.8	0.5	0	0
Crac	0.5455	0.4	1	0.4375	0.2857	0.2	0.4667	1	0.4516	0
Cris	0.5455	0.2	0	0.6875	0.0714	0.9333	0.2	1	0.0323	0
Froo	0.5455	0.2	0.3333	0.3906	0.0714	0.2667	0.8667	0.5	0.0323	0
FroF	0.5455	0	0	0.625	0.0714	0.4667	0.7333	0	0.0161	0
FrMW	0.4545	0.4	0	0	0.2143	0.4667	0.4667	0.5	0.2581	0
FruB	0.6364	0.4	0	0.75	0.3571	0.4667	0.8	1	0.5484	0
JRCN	0.5455	0.2	0.3333	0.5313	0.0714	0.6667	0.4	1	0.129	1
JRFN	0.8182	0.4	0.3333	0.5313	0.1429	0.8667	0.6	1	0.2419	1
MuCB	1	0.4	0.6667	0.4688	0.2143	0.6667	0.8667	1	0.4516	0
Nut&	0.6364	0.2	0.3333	0.5938	0	0.5333	0.6	0.5	0.0645	0
NGAR	0.8182	0.4	0.6667	0.6875	0.2143	0.9333	0.4667	1	0.3548	0
NutW	0.3636	0.4	0	0.5313	0.2143	0.7333	0.1333	1	0.2258	0
Prod	0.4545	0.4	0	1	0.0714	0.8667	0.2	1	0.0806	1
RaBr	0.6364	0.4	0.3333	0.6563	0.3571	0.4667	0.8	0.5	0.7097	0
Rais	0.3636	0.2	0	0	0.1429	0.5333	0.4	1	0.2903	0
RiKr	0.5455	0.2	0	0.9063	0	1	0.2	0	0.0484	0
Smac	0.5455	0.2	0.3333	0.2188	0.0714	0.1333	1	0.5	0.0645	0
Spec	0.5455	1	0	0.7188	0.0714	0.6	0.2	0	0.1129	0

자료 6: 색채환(파일이름: color.txt).

31명의 파장 길이에 따른 14가지 색 인지.

434=남색, 445=청색, 472=청록색, 504=녹색, 555=황녹색, 600=노랑색, 628=귤색, 651=주황색, 674=적색.

434	1													
445	0.86	1												
465	0.42	0.5	1											
472	0.42	0.44	0.81	1										
490	0.18	0.22	0.47	0.54	1									
504	0.06	0.09	0.17	0.25	0.61	1								
537	0.07	0.07	0.1	0.1	0.31	0.62	1							
555	0.04	0.07	0.08	0.09	0.26	0.45	0.73	1						
584	0.02	0.02	0.02	0.02	0.07	0.14	0.22	0.33	1					
600	0.07	0.04	0.01	0.01	0.02	0.08	0.14	0.19	0.58	1				
610	0.09	0.07	0.02	0	0.02	0.02	0.05	0.04	0.37	0.74	1			
628	0.12	0.11	0.01	0.01	0.01	0.02	0.02	0.03	0.27	0.5	0.76	1		
651	0.13	0.13	0.05	0.02	0.02	0.02	0.02	0.02	0.2	0.41	0.62	0.85	1	
674	0.16	0.14	0.03	0.04	0	0.01	0	0.02	0.23	0.28	0.55	0.68	0.76	1

자료 7: 공익회사(파일이름: utility.txt).

22개의 공익회사와 8개 평가변수.

X_1: 고정요금 보상율, X_2: 자금 수익률, X_3: KW당 비용, X_4: 연부하율,
X_5: 1974 ~ 1975년 사이 절정의 kWh 수요성장, X_6: 판매량, X_7: 핵비율,
X_8: 총연료비용.

1.06	9.2	151	54.4	1.6	9077	0	0.628
0.89	10.3	202	57.9	2.2	5088	25.3	1.555
1.43	15.4	113	53.0	3.4	9212	0	1.058
1.02	11.2	168	56.0	0.3	6423	34.3	0.700
1.49	8.8	192	51.2	1.0	3300	15.6	2.044
1.32	13.5	111	60.0	−2.2	11127	22.5	1.241
1.22	12.2	175	67.6	2.2	7642	0	1.652
1.10	9.2	245	57.0	3.3	13082	0	0.309
1.34	13.0	168	60.4	7.2	8406	0	0.862
1.12	12.4	197	53.0	2.7	6455	39.2	0.623
0.75	7.5	173	51.5	6.5	17441	0	0.768
1.13	10.9	178	62.0	3.7	6154	0	1.897
1.15	12.7	199	53.7	6.4	7179	50.2	0.527
1.09	12.0	96	49.8	1.4	9673	0	0.588
0.96	7.6	164	62.2	−0.1	6468	0.9	1.400
1.16	9.9	252	56.0	9.2	15991	0	0.620
0.76	6.4	136	61.9	9.0	5714	8.3	1.920
1.05	12.6	150	56.7	2.7	10140	0	1.108
1.16	11.7	104	54.0	−2.1	13507	0	0.636
1.20	11.8	148	59.9	3.5	7287	41.1	0.702
1.04	8.6	204	61.0	3.5	6650	0	2.116
1.07	9.3	174	54.3	5.9	10093	26.6	1.306

자료 8: 관심영역(파일이름: field.txt).

4개의 여성그룹과 8가지 관심영역. 건강, 과학, 기술, 상업, 사무, 중개, 사회, 예술.

1.00							
0.71	1.00						
0.72	0.67	1.00					
0.60	0.55	0.88	1.00.				
0.55	0.60	0.55	0.55	1.00			
0.63	0.41	0.69	0.71	0.76	1.00		
0.67	0.54	0.61	0.56	0.66	0.76	1.00	
0.63	0.58	0.76	0.76	0.49	0.64	0.59	1.00

1.00							
0.77	1.00						
0.55	0.74	1.00					
0.47	0.61	0.85	1.00.				
0.23	0.23	0.49	0.53	1.00			
0.37	0.36	0.62	0.60	0.74	1.00		
0.59	0.52	0.55	0.49	0.47	0.74	1.00	
0.48	0.59	0.72	0.69	0.43	0.64	0.63	1.00

1.00							
0.80	1.00						
0.54	0.72	1.00					
0.34	0.52	0.84	1.00.				
0.35	0.43	0.66	0.69	1.00			
0.33	0.45	0.62	0.67	0.78	1.00		
0.52	0.65	0.60	0.50	0.54	0.66	1.00	
0.32	0.53	0.79	0.76	0.64	0.72	0.59	1.00

1.00							
0.72	1.00						
0.53	0.69	1.00					
0.45	0.51	0.77	1.00.				
0.17	0.16	0.45	0.40	1.00			
0.31	0.25	0.49	0.38	0.77	1.00		
0.54	0.41	0.47	0.37	0.44	0.67	1.00	
0.38	0.50	0.67	0.55	0.36	0.49	0.55	1.00

자료 9: 국가(파일이름: nation.txt).

2명의 학생(하삼각행렬과 상삼각행렬)과 12개 국가.

Brazil	.	6	8	5	3	4	3	3	2	2	5	2
Congo	7	.	6	7	3	5	4	3	2	2	2	2
Cuba	8	5	.	4	2	3	2	3	3	2	3	2
Egypt	2	7	6	.	2	3	5	2	2	2	2	4
France	3	3	2	6	.	3	4	3	2	4	8	5
India	2	7	2	6	2	.	3	5	5	4	5	4
Israel	3	2	3	8	5	7	.	3	3	4	6	7
Japan	2	1	1	4	3	7	2	.	7	6	7	4
China	2	6	8	7	4	5	3	7	.	8	3	4
USSR	2	3	7	7	3	6	3	6	8	.	5	6
USA	5	3	3	4	8	5	7	8	1	2	.	3
Yugoslavia	2	2	8	3	4	5	4	6	8	8	2	.

자료 10: 모스부호(파일이름: morse.txt).

598명의 아라비아 숫자에 대한 모스부호의 인식

1.---- 2..--- 3...-- 4....- 5..... 6-.... 7--... 8---.. 9----. 0-----

0.000	7.000	11.75	12.69	13.04	12.08	11.36	9.899	7.810	8.602
7.000	0.000	7.550	11.49	12.77	12.12	11.14	11.27	11.14	12.12
11.75	7.550	0.000	9.950	11.05	10.30	11.70	11.92	12.61	12.73
12.69	11.49	9.950	0.000	8.185	10.34	11.22	12.29	12.88	13.00
13.04	12.77	11.05	8.185	0.000	10.77	11.79	12.57	13.08	13.19
12.08	12.12	10.30	10.34	10.77	0.000	6.403	11.40	12.69	12.00
11.36	11.14	11.70	11.22	11.79	6.403	0.000	6.557	10.68	12.21
9.899	11.27	11.92	12.29	12.57	11.40	6.557	0.000	7.937	10.20
7.810	11.14	12.61	12.88	13.08	12.69	10.68	7.937	0.000	5.196
8.602	12.12	12.73	13.00	13.19	12.00	12.21	10.20	5.196	0.000

자료 11: 두개골(파일이름: skull.txt).

40명(22명의 남자(M), 18명의 여자(F)) 두개골과 12개 측정변수.

1) Greatest length, L
2) breadth, B
3) Height, H
4) auricular height, OH
5) Circumference above the superciliary ridger, U
6) Sagittal circumference, S
7) cross-circumference, Q
8) Upper face height, FH
9) Nasal breadth, NB
10) Nasal height, NH
11) cephalic index, B/L
12) Ratio of height to length, H/L

5F	−2.2700	−0.7810	−1.7400	−1.3300	−1.8900	−0.9020	−1.4800	−1.9000	−1.4100	−1.0600	1.4120	0.4960
7M	0.0364	0.3764	−0.9660	−0.6850	0.3166	−0.6520	−0.9220	0.1855	−0.4530	0.4895	0.2161	−1.1200
10F	−0.7070	1.5570	−0.3270	0.1476	−0.4180	−0.4010	1.1240	0.5110	0.6635	−0.8000	1.7180	0.3807
13M	0.9280	1.4390	−0.6750	−0.0960	1.2610	0.1838	1.5570	1.3570	2.5770	0.7475	0.2161	−1.6600
26M	1.6710	1.6760	1.1630	1.6210	2.1710	1.5200	2.0530	0.4459	0.9187	0.2315	−0.2130	−0.4650
32M	−0.5580	0.8488	−1.9300	−1.0700	−0.2780	−1.4000	−0.3640	−0.8560	−0.2930	1.2630	1.0740	−1.5800
43F	−1.4500	0.2583	−1.5500	−1.0100	−1.0100	−1.1500	−0.4880	−0.5960	−1.0900	−0.0260	1.4420	−0.1960
45F	−0.4090	−0.4500	−1.4500	−2.6300	−1.0500	−1.5700	−1.1100	−0.5960	−0.7720	−2.0900	0.0015	−1.2000
46F	−1.3800	1.0850	−0.6750	−0.1730	−0.2080	−0.4010	0.8756	−1.7700	−0.4530	−0.2840	1.9940	0.7268
52M	−0.5580	−0.0960	−0.5790	−0.1730	−0.3130	−0.4010	−0.1160	0.4459	−0.1340	0.7475	0.3694	−0.0810
58F	0.2593	−1.0400	0.1955	0.1476	0.3516	0.1838	−0.2400	1.2270	1.1420	0.7475	−0.9790	−0.0810
59F	−0.1120	−2.0600	−0.7720	0.7241	−0.7330	−0.0670	−0.6120	−1.5100	−1.4100	0.2315	−1.4100	−0.7340
63F	0.5565	0.7307	0.4858	1.2360	0.8414	1.6040	1.2470	−1.2500	−0.7720	0.2315	0.0322	−0.0810
64F	−1.7500	−1.0400	−0.5790	−1.8400	−1.7500	−2.5300	−2.3500	−1.2500	−0.4530	−1.3200	0.5839	1.2650
66M	−0.4090	1.2030	−0.9660	0.0836	−0.0680	−1.4900	0.8756	1.2270	0.6635	2.2950	1.1970	−0.6580
70F	−1.8700	−0.9230	−0.3850	−0.4930	−1.9600	−1.1500	−1.1100	−1.1200	−2.3700	−0.0260	0.9211	1.6110
83F	1.0770	1.5570	−0.3850	0.2117	1.1910	1.1030	0.3797	0.4459	1.3810	−0.8000	0.1854	−1.5000
85M	0.8537	0.8488	0.1955	−0.5570	0.9464	0.3091	−0.3640	−1.6400	−0.7720	−0.8000	−0.1210	−0.6580
86M	−0.1860	0.0221	0.9696	0.3398	−0.0330	0.3091	0.1317	0.7062	−0.2930	−0.8000	0.1548	1.2650
93bM	−1.1500	0.6126	−1.3500	−1.0700	−0.5580	−1.2400	0.2557	−0.4650	0.6635	−1.0600	1.4420	−0.3110
96M	1.0020	1.3210	1.7440	1.7490	1.2610	1.7710	2.1150	−0.5960	−0.4530	0.7475	0.0935	0.8037
97F	−1.6000	1.0850	0.0019	0.1476	−1.3600	−0.7350	−0.1160	0.5761	0.6635	0.8764	2.2090	1.7650
99M	0.9280	0.3764	0.9696	1.3650	0.6665	0.6015	0.7516	−0.2050	−1.2500	−1.0600	−0.5200	0.0731
102M	0.1850	0.1402	0.1955	−0.8130	−0.4880	−0.1500	−0.4880	0.4459	1.1420	−1.3200	−0.0600	−0.0040
112M	0.4822	−2.2200	0.8342	−0.0450	−0.4880	−0.2340	−0.8600	0.6411	1.3020	−0.2840	−1.9600	0.3807
120M	2.1910	0.4945	0.9696	0.2117	2.4510	0.8521	0.5036	1.7480	1.3020	2.2950	−1.3800	−1.1600
121F	0.0364	−0.4500	−1.1600	−1.3300	−0.4180	−0.9020	−0.8600	−0.3350	−0.4530	−0.2840	−0.3660	−1.3100
125M	−0.2610	−0.6860	1.6470	1.2360	−0.3130	0.1838	0.9995	−0.8560	−0.6120	−1.4500	−0.3050	2.0730
136M	1.2250	−1.0400	−0.3850	0.7241	0.8414	1.1030	−0.1160	1.0970	0.5838	−0.0260	−1.7200	−1.6600
137F	0.1850	−0.9230	−0.3850	−0.5570	−0.2080	0.3091	−1.2300	1.4870	0.1053	0.2315	−0.8260	−0.6190
138M	0.3336	−0.3320	−0.1920	0.2117	0.0367	0.1838	−0.1160	−0.8560	−0.7720	0.7475	−0.5200	−0.5810
139M	0.0364	−0.9230	−0.3850	0.0836	−0.6980	−0.0670	−0.3640	−0.3350	0.5040	1.0050	−0.7030	−0.4650
140F	0.7794	0.0221	0.9696	1.7490	0.7714	0.9356	1.6190	0.7062	0.8230	−1.8300	−0.6420	0.2269
143M	0.1850	0.0221	0.7761	−0.4290	0.1067	0.1002	−0.1160	0.3157	−0.2140	−0.5420	−0.1520	0.6499
144F	1.0770	−0.5210	0.9696	1.4930	0.9114	1.1440	0.2557	0.8364	−0.7720	1.0050	−1.2200	−0.0810
145F	0.1850	−0.0960	1.7440	0.7241	0.1417	1.8550	0.5036	−0.0750	0.5040	0.6443	−0.2440	1.6880
146F	0.0364	−0.5090	0.3890	−1.3300	−0.9080	0.1002	−1.4800	0.7062	−0.7720	−0.1550	−0.4280	0.3807
148M	0.1850	−1.7500	0.2922	0.3398	0.0717	−0.1500	−0.4880	−0.8560	−0.7720	−0.0260	−1.4100	0.1115
151M	−1.0000	−0.8050	0.3890	0.0836	−0.6280	−0.1500	−0.8600	−0.0750	0.0255	0.2315	0.2468	1.4960
152M	1.2250	0.9669	1.9370	0.9803	1.4360	1.3950	0.9995	2.0080	1.7800	1.2630	−0.3360	0.8037

자료 12: 유럽 10대 도시(파일이름: eurodist.txt).

유럽 10대 도시 간의 거리

London	Stockholm	Lisboa	Madrid	Paris	Amsterdam	Berlin	Prague	Rome	Dublin
0	569	667	530	141	140	357	396	569	190
569	0	1212	1043	617	446	325	423	787	648
667	1212	0	201	596	768	923	882	714	714
530	1043	201	0	431	608	740	690	516	622
141	617	596	431	0	177	340	337	436	320
140	446	768	608	177	0	218	272	519	302
357	325	923	740	340	218	0	114	472	514
396	423	882	690	337	272	114	0	364	573
569	787	714	516	436	519	472	364	0	755
190	648	714	622	320	302	514	573	755	0

자료 13: 자동차 선호도 평가(파일이름: cartype.txt).

40명의 24대 자동차(모델) 선호도 평가[1(매우 좋다)-6(매우 나쁘다)]의 평균점수와 8개의 평가변수.

X_1: 경제성, X_2: 서비스, X_3: 비감가상각, X_4: 가격(매우 저렴한 차는 1점), X_5: 디자인, X_6: 스포티카 X_7: 안전성, X_8: 다루기 쉬움.

Audi	100	3.9	2.8	2.2	4.2	3.0	3.1	2.4	2.8
BMW	5 series	4.8	1.6	1.9	5.0	2.0	2.5	1.6	2.8
Citroen	AX	3.0	3.8	3.8	2.7	4.0	4.4	4.0	2.6
Ferrai		5.3	2.9	2.2	5.9	1.7	1.1	3.3	4.3
Fiat	Uno	2.1	3.9	4.0	2.6	4.5	4.4	4.4	2.2
Ford	Fiesta	2.3	3.1	3.4	2.6	32	3.3	3.6	2.8
Hyundai		2.5	3.4	3.2	2.2	3.3	3.3	3.3	2.4
Jaguar		4.6	2.4	1.6	5.5	1.3	1.6	2.8	3.6
Lada	Samara	3.2	3.9	4.3	2.0	4.3	4.5	4.7	2.9
Mazda	323	2.6	3.3	3.7	2.8	3.7	3.0	3.7	3.1
Merceedes	200	4.1	1.7	1.8	4.6	2.4	3.2	1.4	2.4
Mitsubishi	Galant	3.2	2.9	3.2	3.5	3.1	3.1	2.9	2.6
Nisan	Sunny	2.6	3.3	3.9	2.1	3.5	3.9	3.8	2.4
Opel	Corsa	2.2	2.4	3.0	2.6	3.2	4.0	2.9	2.4
Opel	Vectra	3.1	2.6	2.3	3.6	2.8	2.9	2.4	2.4
Peugeot	306	2.9	3.5	3.6	2.8	3.2	3.8	3.2	2.6
Renault	19	2.7	3.3	3.4	3.0	3.1	3.4	3.0	2.7
Rover		3.9	2.8	2.6	4.0	2.6	3.0	3.2	3.0
Toyota	Corolla	2.5	2.9	3.4	3.0	3.2	3.1	3.2	2.8
Trabant	601	3.6	4.7	5.5	1.5	4.1	5.8	5.9	3.1
VW	Golf	3.8	2.3	1.9	4.2	3.1	3.6	1.6	2.4
VW	Passat	3.1	2.2	2.1	3.2	3.5	3.5	2.8	1.2
Wartburg	1.3	3.7	4.7	5.5	1.7	4.8	5.2	5.5	4.0

부록 3
프로그램

프로그램 파일(MDS2013.exe: 압축파일)은 아래의 홈페이지를 방문하시어 *Programs in my published books* 메뉴에서 받아 가시기 바랍니다.

home.pusan.ac.kr/~yschoi

[프로그램 압축 풀기]

압축파일 MDS2013.exe을 더블클릭하면 자동으로 MDS2013 폴더아래 프로그램(PGM1-1 ~ PGM7-1, PGM1, PGM2)이 PGM 폴더에 생성되고, [부록 2 : 자료]의 자료파일 13개가 Data 폴더에 생성된다.

[프로그램 목록]

1장. 다차원척도법의 이해

PGM1-1	[프로그램 1.1] 도시 간 철도거리의 계량형 다차원척도법
PGM1-2	[프로그램 1.2] 경제관련기관의 경제전망에 관한 비계량형 다차원척도법 (a) 원자료의 유클리드거리 (b) 이진수자료의 제곱유클리드거리
PGM1-3	[프로그램 1.3] 12개 국가의 유사성행렬자료에 대한 비계량형 다차원척도법

2장. 비유사성과 유사성

PGM2-1	[프로그램 2.1] 원자료로부터 다차원척도법을 위한 SAS/IML과 MDS
PGM2-2	[프로그램 2.2] 이진수자료로부터 다차원척도법을 위한 프로그램
PGM2-3	[프로그램 2.3] 모스부호의 비대칭 유사성자룡의 계량형 다차원척도법을 위한 프로그램

3장. 이원 다차원척도법

PGM3-1	[프로그램 3.1] [알고리즘 2]를 위한 SAS/IML과 PROC GPLOT 프로그램
PGM3-2	[프로그램 3.2] [알고리즘 2]의 비계량형 다차원척도법을 위한 프로그램
PGM3-3	[프로그램 3.3] 모스부호의 유사성자료의 계량형과 비계량형 다차원척도법을 위한 프로그램
PGM3-4	[프로그램 3.4] 모스부호의 비유사성자료의 계량형과 비계량형 다차원척도법을 위한 R-프로그램

4장. 삼원 다차원척도법

PGM4-1	[프로그램 4.1] 삼원 다차원척도법의 대수적 기하적 이해를 위한 예
PGM4-2	[프로그램 4.2] 15군데 신체부위자료에 대한 삼원 다차원척도법

5장. 최대우도 다차원척도법

PGM5-1	[프로그램 5.1] 레크레이션 자료에 대한 최대우도 다차원척도법
PGM5-2	[프로그램 5.2] 레크레이션 자료에 대한 삼원 다차원척도법

6장. 베이즈 다차원척도법

PGM6-1	[프로그램 6.1] 〈표 6.1〉 유럽 10대 도시 간 거리에 대한 베이즈 계량형 다차원척도법을 위한 R-프로그램
PGM6-2	[프로그램 6.2] 〈표 6.1〉 유럽 10대 도시 간 거리에 대한 계량형 다차원척도법을 위한 R-프로그램

7장. 비선형 다차원척도법

PGM7-1	[프로그램 7.1] 모스부호의 비유사성자료의 비선형 계량형 다차원척도법을 위한 R-프로그램

부록 1: SAS MDS 절차

PGM1	[프로그램 1] PROC MDS를 위한 템플릿(template) 절차
PGM2	[프로그램 2] 수정된 PROC TEMPLATE 절차

참고문헌

- 이재창, 박정섭 (1986). 다차원축척(multidimensional scaling)기법, 〈應用統計(고려대학교 통계연구소)〉, 1권 1호, 61-79
- 장광국 (1990). 최소차원분석(MDA)에 의한 자료분석및 다차원축척(MDS)와의 결과 비교. 〈고려대학교 통계학과 석사학위논문〉.
- 정광모, 최용석 (2009). 〈범주형자료분석-SAS의 응용 및 해석-〉, 제2판, 자유아카데미, 서울
- 최병진(1989). 기업의 이미지 유사성과 소비자 선택에 관한 연구, 통계상담 사례집 제1호, 고려대학교 대학원 통계학과.
- 최용석(1995). 〈SAS 다차원척도법〉, 자유아카데미, 서울.
- 최용석(2001). 〈SAS 대응분석의 이해와 응용〉, 자유아카데미, 서울.
- 최용석, 신상민(2013). 〈R을 활용한 행렬도분석의 이해〉, 자유아카데미, 서울.
- 최용석, 정광모(2003). 〈SAS를 활용한 다변량 분석 기법과 응용〉, 자유아카데미, 서울.
- 허명회(1992). 〈수량화 방법론의 이해〉, 자유아카데미, 서울.
- 허명회(1994). 〈SAS 최적척도법: 사례분석을 중심으로〉, 자유아카데미, 서울.
- Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology*, 63, 516-556.
- Borg, I and Groenen, P.J.F. (2005). *Modern Multidimensional Scaling: Theory and Applications*, Springer, New York.
- Boyle, G.J. and Katz, I. (1991). Multidimensional scaling of the eight state questionnaire and the differential emotions scale, *Personality and Individual Differences* 12(6), 565-574.
- Bronstein, A.M., Bronstein, M.M. and Kimmel, R. (2006). Generalized multidimensional scaling: A frame work for isometry-invariant partial surface matching, *Proceedings of the National Academy of Sciences*, 1168-1172.

- Carroll, J.D. and Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Ekart-Young" decomposition, *Psychometrika*, 35, 283-319.
- Catell, R. B.(1966). The scree test for the number of factors. *Journal of Multivariate Behavior Research*, 1, 245-276.
- Chatfield, C. and Collins, A.J. (1980). *Introduction to Multivariate Analysis*, Chapman and Hall, London.
- Choi, Y.S., Jolliffe, I.T. and Lee, J.R. (1998). Robust weighted nonmetric scaling, *The Journal of Data Science and Classification*, 2, 33-47.
- Conte, S.D. and de Boor, C. (1980). *Elementary Numerical Analysis*, McGraw-Hill, New York.
- Cox, T.F. and Cox, M.A.A. (2001). *Multidimensional Scaling*, Chapmann and Hall, London.
- Davison, M.L. (1992). *Multidimensional Scaling*, Krieger, Malabar, Florida.
- Dryden, I.L. and Mardia, K.V. (1998). *Statistical Shape Analysis.* John Wiley & Sons, Chichester.
- du Toit, S.H.C., Steyn, A.G.W. and Stumpf, R.H. (1986). *Graphical Exploratory Data Analysis*, Springer-Verlag, New york.
- Dzwinel, W. (1994). How to make Sammon's mapping useful for multidimensional data structure analysis, *Pattern Recognition*, 27, 949-959.
- Everitt, B.S. and Dunn, G. (1991). *Applied Multivariate Data Analysis*, Edward Arnold, London.
- Forero, P.A. and Giannakis, G.B. (2012). Sparsity-exploiting robust multidimensional scaling, *IEEE Transactions On Signal Processing*, 60(8), 4118-4134.
- Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika*, 53, 325-338.
- Green, P.E. (1975). Marketing Applications of MDS: Assessment and Outlook, *Journal of Marketing*, 39(1), 24-31.
- Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points, *Psychometrika*, 33, 469-506.

- Hordle, W. and Simir, L. (2007). *Applied Multivariate Statistical Analysis*, Springer, Berlin.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, 57, 97-109.
- Horan, C.B. (1969). Multidimensional scaling: Combining observations when individuals have different perceptual structures. *Psychometrika*, 34, 139-165.
- Izenman A.J. (2008). *Modern Multivariate Statistical Techniques*, Springer, New York.
- Jacobowitz, D. (1975). *The Acquisition of Semantic Structure*, Doctorial dissertation, University of North Carolina at Chapel Hill.
- Jobson, J.D. (1992). *Applied Multivariate Data Analysis*, Springer-Verlag, New York.
- Johnson, R.A. and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis*, Prentice-Hall, New Jersey.
- Kennedy, W.J. and Gentle, J.E. (1980). *Statistical Computing*, Marcel Dekker Inc., New York.
- Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, 29, 1-28.
- Kruskal, J.B. and Wish, M. (1978). *Multidimensional Scaling*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-011, Sage Publications, Beverly Hills and London.
- Lawless, H.T., Sheng, N. and Knoops, S.S.C. (1995). Multidimensional scaling of sorting data applied to cheese perception, *Food Quality and Preference*, 6, 91-98.
- Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*, Academic Press, New york.
- Messick, S. J. and Abelson, R. P. (1956). The additive constant problem in Multidimensional scaling. *Psychometrika*, 21, 1-15.
- Oh, M.S. and Raftery, A.E. (2001). Bayesian multidimensional scaling and choice of dimension, *Journal of the American Statistical Association*, 96(455), 1031-1044.
- Okada, K. (2012). A Bayesian approach to asymmetric multidimensional scaling.

Behaviormetrika, 39, 1-14.

- Okada, K. and Mayekawa, S. (2011). Bayesian nonmetric successive categories multidimensional scaling. *Behaviormetrika*, 38, 17-31.
- Okada, K. and Shigemasu, K. (2009). BMDS: A collection of R functions for Bayesian Multidimensional Scaling. *Applied Psychological Measurement*, 33, 570-571.
- Okada, K. and Shigemasu, K. (2010). Bayesian multidimensional scaling for the estimation of a Minkowski exponent, *Behavior Research Methods*, 42, 899-905.
- Ramsay, J.O. (1977). Maximum likelihood estimation in multidimensional scaling, *Psychometrika*, 42, 241-266.
- Ramsay, J.O. (1980). Some small sample results for maximum likelihood estimation in multidimensional scaling, *Psychometrika*, 45, 139-144.
- Ramsay, J.O. (1982). Some statistical approaches to multidimensional scaling data, *Journal of Royal Statistical Society,* A(145), 285-312.
- Ramsay, J. O.(1983). *Multiscale Manual*, Scientific Software, Chicago. Rider, D. de and Duin, R.P.W. (1997). Sammon' s mapping using neural networks: A comparison, *Pattern Recognition Letters,* 18, 1307-1316.
- Sammon, J.W. (1969). A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers*, C-18, 401-409.
- SAS Institute Inc. (1985). *SAS Language Guide for Personal Computers,* Version 6, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1992). *SAS Technical Report p-229 SAS/STAT Software; Changes and Enhancements*, Release 6.07. Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (2004). *SAS/IML 9.1 User's Guide,* SAS Institute Inc., Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (2008). *SAS/STAT 9.2 User's Guide,* SAS Institute Inc., Cary, NC: SAS Institute Inc.
- Schiffman, S.S., Reynolds, M.L. and Young, F.W. (1981). *Introduction To Multidimensional scaling*, Academic, New York.
- Seber, G.AF.(1984). Multivariate Observations, John Wiley & Sons, Inc., New York.
- Shepard, R.N. (1962). The analysis of proximities; Multidimensional scaling with

an unknown distance function. Ⅰ and Ⅱ, *Psychometrika*, 27, 125-140, 219-246.
- SPSS Inc.(1992) SPSS for Windows: Professional Statistics, Release 5, 173.
- Stovel, K., Savage, M. and Bearman P. (1996). Ascription into achievement: Models of career systems at Lloyds Bank, 1890-1970, *American Journal of Sociology*, 102(2), 358-399.
- Takane, Y., Young, F.W. and de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features, *Psychometrika*, 42, 7-67.
- Torgerson, W. S. (1952). Multidimensional scaling: Ⅰ. Theory and Method. *Psychometrika*, 17, 401-419.
- Torgerson, W. S.(1958). *Theory and Methods of Scaling.* Wiley, New York.
- Torgerson, W. S. and Meuser, G. (1962). Informal notes on Torgerson and Meuser's IBM 7090 program for multidimensional scaling. *Mimeographed report*, Johns Hopkins University, Baltimore, MD.
- Tucker, L.R. and Messick, S. (1963). An individual differences model for multidimensional scaling, *Psychometrika*, 28, 333-367.
- Young, F.W. (1987). *Multidimensional Scaling: History, Theory, and Applications* (edited by Hamer, R.M.), NJ: Lawrence Erlbaum Associates, Hillsdale.
- Young, F.W. and Lewyckyj, R. (1979). *ALSCAL-4 User's Guide, 2nd edition*, *Data Analysis and Theory Associates*.
- Young, G. and Householder, A.S. (1938). Discussion of a set of points in terms of their mutual distances, *Psychometrika*, 3, 19-22.

찾아보기

[ㄱ]
가우스-뉴튼법 • 24, 28
가중 다차원척도법 • 3, 102
가중유클리드거리 44, 102, 120, 116
가중치 • 19, 149, 170
가중치행렬 • 102, 120, 131
결합사후밀도함수 • 140, 141, 150
개인지각공간 • 103, 104, 120, 130
개체형상좌표 • 105
결측치 • 168, 176, 179
계량형 다차원척도법 • 2, 18
계량형 해 • 68
고전적척도법 • 2
곱적률 상관계수 • 50, 52, 63, 64
공간근사 • 21, 138, 150, 158, 164
교대최소제곱법 • 131, 172
근접성 • 40

[ㄷ]
다차원척도그림 • 1, 2
다차원척도법 • 1, 2, 15
단순매칭계수 • 8, 29
단조 • 21, 74, 101, 172
단조함수 • 21, 24, 101, 121, 173
도식도 • 101, 102
독립성 검정 • 52
대각계량형모형 • 123
등분산 • 122, 124

[ㄹ]
라그랑지승수 • 123
러셀-라오계수 • 48, 63, 64, 91
로그우도함수 • 122, 123
로그-정규분포 • 119, 121, 131

[ㅁ]
마할라노비스거리 • 42, 94, 97, 166
맨하턴거리 • 41
몬테칼로방법 • 138
멱승변환 • 129
메트로폴리스-해스팅스 • 138, 142
민코브스키거리 • 41

[ㅂ]
반대칭행렬 • 57, 58, 59, 65
벌점 • 145
범주형자료 • 42, 43, 52
범주형자료분석 • 52
불편추정치 • 123
베이즈 다차원척도법 • 3, 19
베이즈추정치 • 123
비관측 • 139
비계량형 다차원척도법 • 2, 18

비선형 다차원척도법 • 4, 19, 157, 198
비선형 모형 • 158
비선형회귀 • 24
비선형사상 알고리즘 • 4, 19, 157, 158, 159, 160, 164, 165
비유사성 • 3, 4
비유사성행렬 • 2, 20, 92, 140, 161, 162, 166, 168, 170, 171

[ㅅ]
사전분포 • 140, 148
삼원 다차원척도법 • 99
상그림 • 77, 82, 112, 127, 135, 148, 168, 171, 177
세퍼드그림 • 77, 83, 114, 167, 168, 171
새먼사상 • 4, 157, 158
새먼의 스트레스 • 158, 159, 164
수치적분 • 141
순서척도 • 2
순위상 • 75
스트레스 • 22
스트레스 공식 1과 2 • 23
S-스트레스 공식 1과 2 • 23
스트레스-차원수 그림 • 24, 33, 69, 77, 116, 136, 168, 171
스트레스 최소화 알고리즘 • 4, 23
스펙트럼분해 • 69, 70
시티-블럭거리 • 18

[ㅇ]
양반정치 • 49, 53
양적 • 39, 43, 44, 50, 51, 77
역-감마분포 • 140, 142
연관성 • 8, 40, 64, 86, 165
완전조건부사후밀도함수 • 140
우도함수 • 122, 123, 130
유사성계수 • 39, 48, 49, 51
유클리드거리 • 2
이상치 • 19, 77, 82, 149, 156
이원 다차원척도법 • 67
이중-중심화행렬 • 68
이진수 • 6, 29, 50, 63, 74, 91, 152, 196
인자분석 • 24, 26, 27, 70
임의보행 • 138, 142

[ㅈ]
잔차제곱합 • 22, 139, 141, 145
적합도 • 26, 69
절단 정규분포 • 139
정규경사하강법 • 160
정방대칭행렬 • 103
조합표 • 57, 58, 59, 60
주변사후분포함수 • 144
주성분분석 • 26, 36, 68, 70, 87, 157, 164
주성분좌표분석 • 68
중요가중치 • 100
중요점가중치 • 100
질적 • 39, 51
짝특수분산 • 121, 122

[ㅊ]
차원가중치 • 100
차원계수 • 104
차원수 • 26
척도변환 • 22, 129
초계량형 • 40
초깃값 • 90, 141
측정수준 • 10, 20, 32, 34, 35, 69, 138, 169, 177

측정척도 • 2, 20
척도인자 • 22, 23
최대우도 다차원척도법 • 120, 130, 169, 171, 197
최대우도 추정법 • 119
최적적합그림 • 148, 149

[ㅋ]
카이제곱통계량 • 52
크루스칼-세퍼드 알고리즘 • 18, 75, 83
크루스칼의 판별기준 • 26, 69, 77, 86, 113, 130, 164

[ㅌ]
토거선 알고리즘 • 68, 70, 83, 87, 92, 153, 164
템플릿 • 177, 178, 198

[ㅍ]
표준변환 • 53, 59. 85, 88, 95, 153, 161
피어슨거리 • 41, 42, 166
프로크러스티즈 • 142, 143

[ㅎ]
항등계량형 • 121
형상공간 • 2, 69
형상좌표 • 28, 46
형상좌표행렬 • 82
행렬분해 • 58, 59

[A]
ALSCAL • 3
ALSOS 173
ALTERNATE= • 169, 173, 175
Alternating least square • 19, 109
Asymmetric • 56, 59, 60
Association • 40, 48
Association table • 48

[B]
Badness-of-Fit Criterion • 28, 46, 79, 83, 113, 130, 171
Bayesian • 3, 19
Bayesian MDS • 3, 19, 137
Binary • 6, 54
Biplot • 87
bmds • 146, 147, 148

[C]
cmdscale • 88, 89, 151, 160, 162
City-block distance • 18
Classical scaling • 2
Closeness • 40
COEF= • 44, 110, 126, 131, 170
CONDITION= • 110, 126, 131, 170, 171
Conditional moment estimate • 122
Configuration space • 2
Conjugate • 140
CONVERGE= • 29, 31, 47, 56, 169, 175
Correspondence analysis • 43
CUTOFF= • 169, 176, 179

[D]
DECIMALS= • 169

Diagonal metric • 121
DIMENSION= • 29
Dimension coefficient • 104
Dissimilarity • 2, 40
Dissimilarity matrix • 2

[E]
EPSILON= • 169, 176
Euclidean distance • 2, 46, 47

[F]
FIT= • 29
FORMULA • 20
FORTRAN • 3, 138, 146

[G]
Gauss-Newton method • 24
GCONVERGE= • 29
Goodness of fit • 26
Group stimulus • 103

[I]
Identity metric • 121
Image diagram • 77
Importance • 100
INDSCAL(individual differences scaling) • 3
INITIAL= • 159, 161, 164
INAV= • 169, 176
isoMDS • 88, 89, 157, 162

[L]
LEVEL= • 28
Levenberg-Marquardt • 24, 28
Log likelihood function • 122
Lognormal distribution • 121

[M]
Mahalanobis distance • 46
Markov chain Monte Carlo • 137
MASS • 21, 88, 89, 162
Maximum likelihood estimation method • 3
Maximum likelihood MDS • 119
MAXITER= • 29
MCONVERGE= • 31
MCMC • 3, 19
MDS • 1, 3
MDSIC • 147, 148
Metric MDS • 60, 71
Metropolis-Hastings • 138
MINCRIT • 179, 175, 179
Missing value • 170
Monotone • 31, 47, 56, 75, 175

[N]
NEGATIVE • 169, 176, 178, 179
Nonlinear MDS • 4, 19, 157
Nonlinear regression • 24
NONORM • 169, 176
NOPHIST • 24
Normal gradient descent • 160
NOULB • 169, 179
Numerical integration • 141

[O]
OCOEF • 169, 171, 172
OCONFIG • 172
OCRIT • 24, 25, 169, 171, 172
ODS GRAPHICS ON • 28
OTRANS • 169, 171, 172
OUT= • 24, 114
OUTFIT= • 169, 171
OUTITER • 169, 171
Outlier • 77
OUTRES • 79, 114, 169, 171
OVER • 29, 31, 47, 169, 175, 178

[P]
Pair specific component • 121
PCOEF • 169, 174
PCONFIG • 28, 30, 169, 174
PDATA • 169, 174
Penalty • 145
PFINAL • 29, 46, 79
PFIT • 169, 174
PFITROW • 169, 174
PINEIGVAL • 169, 174
PININ • 169, 174
PITER • 169, 174, 181
Post-processing • 142
Positive semidefinite • 53
Principal coordinate analysis • 68
Prior • 140
Private perceptual space • 103
PROC ALSCAL • 19
PROC MDS • 1, 4, 19
PROC MLSCALE • 180, 181
PROC TEMPLATE • 168, 177, 178, 198
Product moment correlation • 50
Procrustes • 142
Proximity • 40
PTRANS • 169, 174

[Q]
Quadratic • 140
Quantative • 43
Qualitative • 43

[R]
Rank image • 75, 76
Residual sum of squares • 22
RIDGE= • 29, 31, 47, 56, 169, 175
RANDOM • 138, 169, 176
Random walk • 138
RSS • 60, 61

[S]
Salience • 100
Sammon mapping • 4, 157
SAS/IML • 4, 44
Scale factor • 4, 19, 105, 167
Schematic diagram • 102
Scree plot • 24
SHAPE= • 29, 31, 56, 126
Simple matching coefficient • 54
SIMILAR= • 34, 169, 170
Similarity coefficient • 48
SINGULAR= • 169, 176
Skew • 57
Skew-symmetric • 59, 60, 61
Spectral decomposition • 71, 74
Stress • 22, 25, 157, 162
Subject • 3, 13, 99

Subject specific variance • 121

[T]

Transformation • 39, 53, 60, 85, 86
Truncated • 139, 150
Two-way MDS • 67

[U]

Ultrametric • 40
UNITE • 16
Unobserved • 139

[W]

Weight • 3, 19
Weighted MDS • 102
within-subject variance • 122
within-pair variance factor • 122

| 저자 소개 |

최 용 석

부산대학교 계산통계학과 졸업
고려대학교 대학원 통계학과 석, 박사
영국 애버딘대학 수리과학과(통계전공) Post-Doc.
현) 부산대학교 통계학과 교수
E-mail: yschoi@pusan.ac.kr

주요 저서) 「R을 활용한 행렬도 분석의 이해」
「통계상담」
「SAS를 활용한 다변량 분석기법과 응용」
「범주형 자료분석 -SAS의 응용 및 해석-」

다차원척도법의 산책

| 저 자 최용석
| 발 행 인 주정희
| 발 행 처 자유아카데미
| 주 소 경기도 파주시 회동길 37-42
파주출판도시
| 전 화 031-955-1321
| F A X 031-955-1322
| 홈페이지 www.freeaca.com
| 전자우편 main@freeaca.com(대표)
editor@freeaca.com(편집)
| 등 록 제406-2003-017호, 1980. 7. 12
| 제1판1쇄 2014년 2월 20일 발행
| 정 가 원

저자와의
협의하에
인지생략

무단 복제를 금하며, 잘못된 책은 바꾸어 드립니다.
ISBN 978-89-7338-