The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.4.613
25(4), 613–620

Multidimensional Scaling of Asymmetric Distance Matrices

Myung-Hoe Huh[1] · Yonggoo Lee[2]

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

(Received February 29, 2012; Revised May 14, 2012; Accepted June 13, 2012)

Abstract

In most cases of multidimensional scaling(MDS), the distances or dissimilarities among units are assumed to be symmetric. Thus, it is not an easy task to deal with asymmetric distances. Asymmetric MDS developed so far face difficulties in the interpretation of results. This study proposes a much simpler asymmetric MDS, that utilizes the notion of "altitude". The analogy arises in mountaineering: It is easier (more difficult) to move from the higher (lower) point to the lower (higher). The idea is formulated as a quantification problem, in which the disparity of distances is maximally related to the altitude difference. The proposed method is demonstrated in three examples, in which the altitudes are visualized by rainbow colors to ease the interpretability of users.

Keywords: Multidimensional scaling(MDS), similarity, asymmetric distance matrix, altitude model, social network analysis.

1. 연구배경과 목적

다차원 척도법(multidimensional scaling), 일명 MDS는 n개 개체 간 유사도(similarity) 또는 비유사도(dissimilarity)가 관측된 자료를 저차원 공간에 타점하는 기법이다. Torgerson의 고전적(classical) 계량형 MDS부터 비계량형(nonmetric) MDS까지 여러 종류가 개발되어 있다. 통상적인 MDS 기법들은 $n \times n$ 대칭적 거리행렬을 대상으로 한다.

Table 1.1은 8종의 심리학 저널 간 상호인용 자료로서 (Lattin 등, 2003), 칸 (i, j)의 빈도 a_{ij}는 저널 i의 논문에 인용된 저널 j의 논문 수이다. 따라서 a_{ij}는 일종의 유사도이지만 $a_{ij} \neq a_{ji}$이다. 이와 같은 상황에서 비대칭적인 $\mathbf{A}\,(= (a_{ij}))$를 대칭적인

$$\frac{1}{2}(\mathbf{A} + \mathbf{A}^t)$$

로 대체한 다음 MDS를 적용하게 되는데, 이렇게 하면 자료의 기본특성이 유실될 수밖에 없다.

비대칭적 MDS의 효시는 Young (1975)에 의해 제안된 `ASYMSCAL`이다. `ASYMSCAL`에서는 개체 i로부터

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

Table 1.1. Cross-References among eight psychology journals (Lattin, *et al.*, 2003)

	AJP	JAS	JAP	CPP	JCP	JED	JEX	PKA	합계
AJP	119	32	2	35	6	4	125	2	325
JAS	8	510	8	8	116	9	19	5	683
JAP	4	16	84	0	11	7	6	5	133
CPP	21	11	1	533	1	0	70	0	637
JCP	0	73	7	0	225	3	0	13	321
JED	1	9	8	1	7	52	0	2	80
JEX	85	119	16	126	12	27	586	13	984
PKA	2	4	10	1	7	5	15	58	102
합계	240	774	136	704	385	107	821	98	3265

AJP: American Journal of Psychology, JAS: Journal of Abnormal and Social Psychology, JAP: Journal of Applied Psychology, CPP: Journal of Comparative and Physiological Psychology, JCP: Journal of Consulting Psychology, JED: Journal of Educational Psychology, JEX: Journal of Experimental Psychology, PKA: Psychometrika.

개체 j에 이르는 거리(비유사도) d_{ij}를

$$d_{ij} = \sum_{k=1}^{s} w_{ik}(x_{ik} - x_{jk})^2$$

으로 모형화하는데, 여기서 $\mathbf{x}_i = (x_{i1}, \ldots, x_{is})$는 s-차원 MDS 공간에서 개체 i의 타점 좌표이고 w_{ik}는 개체 i가 MDS 공간의 k-축에 부여한 가중치이다. 따라서 `ASYMSCAL`은 개인 차가 고려된 Carroll과 Chang (1970)의 `INDSCAL`과 같은 선상에 있는데 개체 별로 각 차원에 부여하는 가중치가 다르므로 출력결과의 시각적 표현이 어렵다. 그런 이유로, 현재까지도 `ASYMSCAL`은 일반 분석자들에 보급되지 않고 있다.

그 이후 비대칭적 MDS 방법으로 다수의 알고리즘이 제안되었다 (Cox와 Cox, 2001; Chino, 2006; Chino, 2011). 그 중 대표적인 Constantine과 Gower (1978)의 방법은 비대칭 행렬 $\mathbf{A}$를 symmetric 부분 $(1/2)(\mathbf{A} + \mathbf{A}^t)$와 skew-symmetric 부분 $(1/2)(\mathbf{A} - \mathbf{A}^t)$로 분리하여 전자로부터 통상적인 MDS 그래프를 만들고 후자로부터는 특이값 분해(singular value decomposition)에 의한 저차원 시각화를 만들어낸다. 이에 따라 2개의 그래프가 연계되어 해석되어야 하는데 그렇게 하더라도 비대칭성의 방향은 그래프로부터 알아내기 어렵다.

본 연구의 목적은 비대칭 거리행렬에 대한 해석이 쉽고 명료한 시각적 MDS 플롯을 제시하는 데 있다. 이 방법은 대칭화 거리행렬에 대한 MDS 플롯에 n개 개체 간 거리의 비대칭성을 '고도(高度, altitude)'로 변환하여 컬러 스펙트럼으로 표현하므로 일반 분석자들이 비대칭적 MDS 결과를 쉽고 명확하게 이해할 수 있다.

2절에서 이 방법을 제안하고, 3절에서 8종 심리학 저널의 상호인용 자료와 EIES 연구자 그룹 32명 연구자 간 메일발송 사례, 그리고 Knoke의 정보 네트워크 사례에 이 방법을 적용하여 보인다.

2. 제안 방법

비대칭 거리행렬 $\mathbf{D}\,(= (d_{ij}))$가 분석자료라고 하자. 만약 원자료가 유사성 측도로 얻어졌다면 적절한 변환을 거쳐서 거리(비유사성)으로 재표현되었음을 가정한다.

Table 2.1. Hiking time of the Mt. Buk Han "U Ee Dong ↔ Baek Woon Dae" course

	A	B	C	D	E
A: U Ee Dong	0	40	65	120	140
B: Do Seon Sa	30	0	25	80	100
C: Ha Ru Jae	50	20	0	55	75
D: Wee Moon	90	60	40	0	20
E: Baek Woon Dae	105	75	55	15	0

$\mathbf{D}^t$와 $\mathbf{D}$의 차이를 $\mathbf{E}\,(=(e_{ij}))$로 정의하자. 즉,

$$e_{ij} = d_{ji} - d_{ij}.$$

따라서 $e_{ij} + e_{ji} = 0$이다. $e_{ij} > 0$이라는 것은 $d_{ij} < d_{ji}$임을, 즉 산행(山行)에서 i가 j보다 높은 곳에 위치함을 뜻한다. 왜냐하면 i가 j보다 높은 곳에 있는 경우 i로부터 j에 가는 것이 j로부터 i로 가는 것보다 쉽기 때문이다. 이런 생각을 살려 개체들의 높낮이를 정하기로 한다.

개체 $i\,(=1,\ldots,n)$에 고도(高度, altitude) z_i를 부여하자. $e_{ij} > 0$인 경우엔 $z_i > z_j$가 되게 하고 반대로 $e_{ij} < 0$인 경우엔 $z_i < z_j$가 되도록, $z_1,\ldots,z_n$을 만들 수 있을까? 이를 위해서, 다음 정식화를 고려한다.

$$\max_{z_1,\ldots,z_n} \sum_{i=1}^{n}\sum_{j=1}^{n} e_{ij}(z_i - z_j), \quad \text{subject to} \quad \sum_{i=1}^{n} z_i^2 = 1 \quad \text{and} \quad \sum_{i=1}^{n} z_i = 0. \tag{2.1}$$

그 해는 다음과 같이 얻어진다.

함수 $\phi(z_1,\ldots,z_n)$을

$$\phi = \sum_{i=1}^{n}\sum_{j=1}^{n} e_{ij}(z_i - z_j) - \lambda \sum_{i=1}^{n} z_i^2$$

으로 놓자. 그러면

$$\frac{\partial \phi}{\partial z_i} = \sum_{j=1}^{n} e_{ij} - \sum_{j=1}^{n} e_{ji} - 2\,\lambda z_i = 2\sum_{j=1}^{n} e_{ij} - 2\,\lambda z_i$$

이므로

$$\frac{\partial \phi}{\partial z_i} = 0 \quad \Rightarrow \quad z_i = \frac{1}{\lambda}\sum_{j=1}^{n} e_{ij} \;\propto\; e_{i+}.$$

따라서 $\mathbf{z}\,(=(z_1,\ldots,z_n))$의 해는

$$\mathbf{z} = \frac{\mathbf{e}^{[1]}}{\|\mathbf{e}^{[1]}\|}$$

이다 ($\sum_{i=1}^{n} z_i = 0$, $\sum_{i=1}^{n} z_i^2 = 1$). 여기서 $\mathbf{e}^{[1]} = (e_{1+},\ldots,e_{n+})$, $e_{i+} = \sum_{j=1}^{n} e_{ij}$이다. 수치 예로서, 고도가 실측된 한 사례에 이상의 수량화를 적용해보기로 하자. Table 2.1은 북한산 등산로 "우이동 ↔ 백운대" 코스의 5개 지점의 "행 → 열" 간 소요시간이다 (단위: 분).

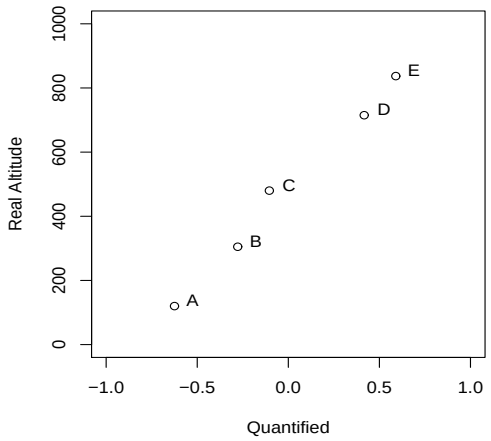

Figure 2.1. Quantified vs. Measured Altitudes of the Mt. Buk Han hiking course

앞의 알고리즘에 따라, 5개 지점(A: 우이동, B: 도선사, C: 하루재, D: 위문, E: 백운대)의 고도를 수량화한 결과는 다음과 같다:

$$-0.62, \quad -0.28, \quad -0.10, \quad +0.42, \quad +0.59.$$

한편, 5개 지점의 실측고도는 다음과 같다 (단위: meter):

$$120, \quad 305, \quad 480, \quad 715, \quad 837.$$

Figure 2.1은 북한산 등산로 5개 지점의 수량화된 고도와 실측고도의 연관성을 보여주는 산점도인데 두 변수 간 선형성은 거의 완벽해 보인다. 이와 같이, 지점 간 이동시간의 차이 즉 비대칭성만으로도 지점 간 상대적 고도의 차이를 유추할 수 있다.

3. 사례분석

심리학 저널의 상호인용 사례

Table 1.1의 8종 심리학 저널의 상호인용 빈도 a_{ij}는 저널 i의 논문에 인용된 저널 j의 논문 수이므로 일종의 유사도로 볼 수 있다. 이것을 거리로 하기 위해서 다음과 같이 역수 변환을 취하기로 한다.

$$d_{ij} = \frac{1}{1 + a_{ji}}.$$

예컨대, AJP로부터 JAS에 이르는 거리는 $1/(1+8)$이고 JAS로부터 AJP에 이르는 거리는 $1/(1+32)$이다. 이처럼 JAS로부터 AJP에 이르는 거리가 AJP로부터 JAS에 이르는 거리보다 짧은 이유는 JAS의 논문이 AJP 논문에 인용된 횟수가 AJP 논문이 JAS 논문에 인용된 횟수보다 크기 때문이다. 이에 따라 $a_{ij} > a_{ji}$이면 $d_{ji} < d_{ij}$가 되고 $e_{ij} = d_{ji} - d_{ij} < 0$이 된다. 즉 저널 i가 저널 j를 끌어오는 빈도가 역의 빈도보다 큰 경우에는 z_i가 z_j보다 작게 된다.

다음은 2절에서 제안된 수량화 방법으로 8종 저널에 부여된 '고도'이다.

저널:	AJP	JAS	JAP	CPP	JCP	JED	JEX	PKA
고도:	− 0.32	0.05	0.13	0.00	0.68	0.23	−0.57	−0.19

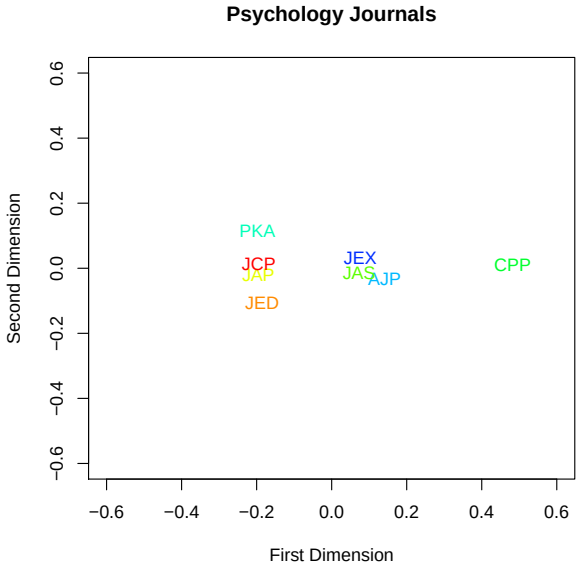

Figure 3.1. Metric MDS of the Cross-Reference data

JEX가 가장 낮고 이어서 AJP가 낮으며 반대로 JCP가 가장 높고 그 다음이 JED인 것으로 나타났다. 이것은 JEX가 타 학술지 논문을 끌어 쓰는 경향이 있음을 의미한다. JCP는 JEX와 대비되는 성격의 학술지인 듯하다. Table 1.1을 보면, JEX 논문이 JCP 논문을 인용한 횟수가 12이나 JCP 논문이 JEX 논문을 인용한 횟수는 0이다.

Figure 3.1는 거리 행렬 $\mathbf{D}$와 $\mathbf{D}^t$를 조화 평균하여 생성된 대칭화 거리 행렬에 적용하여 얻은 계량형 MDS 플롯이다 (`R stats`의 `cmdscale` 함수 사용). 각 개체에 부여된 색은 고도를 나타낸다. 실제로는 고도의 순위(rank)에 따라 무지개 스펙트럼의 한 색상이 부여되었다. 빨강-주황-노랑-초록-파랑의 순서로 (JCP-JED-JAP-JAS-CPP-PKA-AJP-JEX의 순서로) 고도가 높고 낮다.

식 (2.1)의 최적화로부터 $\mathbf{E}\,(=(e_{ij}))$와 $\mathbf{Z}\,(=(z_i - z_j))$간 상관(correlation)이 고도의 수량화에서 적합도임을 알 수 있다. 그 값을 산출한 결과는 0.65이다.

EIES 연구자 그룹 사례

Freeman과 Freeman (1980)의 EIES 연구 net.3 자료는 32명의 연구자 그룹에서 발송 전자우편 발송 도수를 담고 있다 (이에 대한 설명은 Wasserman과 Faust (1994) 참조). 즉 32×32 행렬 $\mathbf{A}$의 칸 요소 a_{ij}는 연구자 i가 연구자 j에 발송한 전자우편 수이다. 32명의 총 992개$(= 32 * 31)$ 순서쌍 중에서 552개 순서쌍에서 a_{ij}가 0이다. i로부터 j에 이르는 거리를 다음과 같이 정의하기로 한다.

$$d_{ij} = \frac{1}{1 + a_{ij}}.$$

이에 따라 $a_{ij} > a_{ji}$이면 $d_{ij} < d_{ji}$가 된다. 즉 i가 j로 보낸 메시지 수가 역방향 메시지 수보다 큰 경우에는 z_i가 z_j보다 크게 된다.

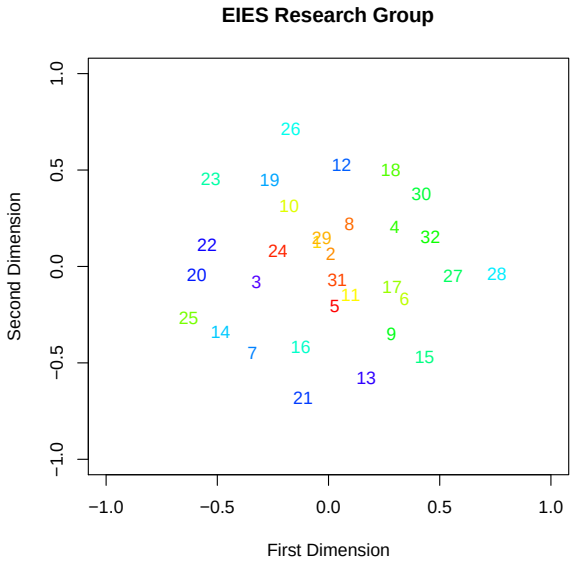

Figure 3.2. Nonmetric MDS of the EIES research group

다음이 2절의 수량화 방법으로 32명 연구자에 부여된 '고도'이다.

연구자 :	3	13	22	20	21	⋯	2	8	31	24	5
고도 :	− 0.24	− 0.19	− 0.19	− 0.19	− 0.18	⋯	0.16	0.16	0.18	0.50	0.55

Figure 3.2은 거리 행렬 $\mathbf{D}$와 $\mathbf{D}^t$를 조화 평균하여 생성된 대칭화 거리 행렬에 적용하여 얻은 비계량형(nonmetric) MDS 플롯이다 (`R MASS`의 `isoMDS` 함수 사용, Venables와 Ripley, 2002). 각 연구자는 고도의 순위에 대응하는 색으로 표현되어 있다. 빨강-주황-노랑-초록-파랑의 순서로 고도가 높고 낮다. 빨강으로 색칠 된 5, 24, 31, 8, 2 등이 가운데에 모여 있음을 볼 수 있는데 이는 전자우편 다(多) 발송자들이 연구자 그룹에서 중심적 역할을 하고 있음을 의미한다. 고도 수량화의 적합도를 나타내는 $\mathbf{E}$와 $\mathbf{Z}$ 간 상관계수는 0.62이다.

Knoke의 정보 네트워크 사례

Knoke의 정보 네트워크(information network)는 미국의 한 중서부 도시의 사회복지 기관 간 정보의 흐름을 나타낸다 (Hanneman and Riddle, 2005). 이 사례에서 기관 리스트는 [1] COUN, [2] COMM, [3] EDUC, [4] INDU, [5] MAYR, [6] WRO, [7] NEWS, [8] WAY, [9] WELF, [10] WEST 등 10개이다.

Table 3.1에 제시된 기관 간 최단거리 행렬에서 INDU → COUN의 거리는 1인데 COUN → INDU의 거리는 2이다. 이와 같은 비대칭성은 네트워크가 방향성을 갖기 때문이다.

10개 기관의 고도를 구한 결과는 다음과 같고 Figure 3.3은 이것이 표현된 비계량 MDS 플롯이다 (`R MASS`의 `isoMDS` 함수 사용). 이 그림에서는 정보 전달적 관계가 연결선으로 표시되어 있다.

기관:	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
고도:	− 0.09	− 0.09	0.18	− 0.18	0.00	0.46	− 0.64	0.37	− 0.28	0.28

Table 3.1. Shortest distances in the Knoke's information network (Hanneman and Riddle, 2005)

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
COUN	0	1	2	2	1	3	1	2	1	2
COMM	1	0	1	1	1	2	1	1	1	2
EDUC	2	1	0	1	1	1	1	2	2	1
INDU	1	1	2	0	1	3	1	2	2	2
MAYR	1	1	1	1	0	2	1	1	1	1
WRO	3	2	1	2	2	0	1	3	1	2
NEWS	2	1	2	1	1	3	0	2	2	2
UWAY	1	1	2	1	1	3	1	0	1	2
WELF	2	1	2	2	1	3	1	2	0	2
WEST	1	1	1	2	1	2	1	2	2	0

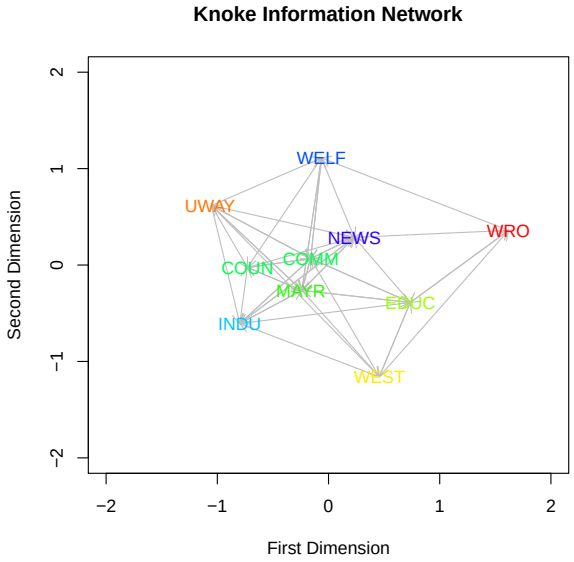

Figure 3.3. Nonmetric MDS of the Knoke information network

NEWS(신문사)의 고도가 –0.64로 가장 작다. 이는 NEWS가 가장 강한 정보흡수 노드임을 의미한다. WELF와 INDU는 NEWS 다음으로 강한 정보흡수 노드이다. 반면, WRO(세계구호기구)의 고도가 +0.46으로 가장 크다. 이는 WRO가 가장 강한 정보발생 노드임을 의미한다. 그 외, UWAY, WEST도 WRO 다음으로 강한 정보발생 노드이다. 연결정도(degree)가 큰 MAYR(市長) 노드와 COMM(상공회의소) 노드의 고도는 0에 가깝다. 두 노드는 정보발생과 정보흡수의 강도가 비슷하다. 고도 수량화의 적합도를 나타내는 **E**와 **Z** 간 상관계수는 0.73이다.

4. 맺음말

비대칭적 MDS에 대하여는 Young (1975) 이래 여러 연구가 있었다. 본 논문에서 제안된 고도 모형과 컬러링은 쉽고 명료한 해석이 최대의 장점이다. 이 연구의 고도모형은 2개 개체 간 차이 $d_{ij} - d_{ji}$를 고

도의 차이 $z_i - z_j$에 대응시키는 선형적 수량화 방법인데, 향후 연구에서 $d_{ij} - d_{ji}$가 $z_i - z_j$의 단조적 증가 함수로 최적 근사되는 비선형적 알고리즘이 개발되길 기대한다.

References

Carroll, D. J. and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-Way generalization of "Eckart-Young" decomposition, *Psychometrika*, **35**, 283–319.

Cox, T. F. and Cox, M. A. A. (2001). *Multidimensional Scaling*, Second Edition, Chapman & Hall/CRC, Boca Raton, Section 4.8.

Chino, N. (2006). Asymmetric multidimensional scaling and related topics, *Manuscript for the invited talk at the Weierstrass Institute for Applied Analysis and Stochastics*, Berlin.

Chino, N. (2011). Asymmetric multidimensional scaling: 1. Introduction, *Journal of the Institute for Psychological and Physical Sciences*, **3**, 101–107.

Constantine, A. G. and Gower, J. C. (1978). Graphical representation of asymmetry, *Applied Statistics*, **27**, 297–304.

Freeman, L. C. and Freeman, S. C. (1980). A semi-visible college: Structural effects on a social networks group. 77–85 in Henderson, M. M. and McNaughton, M. J. (eds.) *Electronic Communication: Technology and Impacts*, Boulder, CO: Westview Press.

Hanneman, R. A. and Riddle, M. (2005). *Introduction to Social Network Methods*, Available at http://www.faculty.ucr.edu/∼hanneman/nettext/.

Lattin, J., Carroll, J. D. and Green, P. E. (2003). *Analyzing Multivariate Data*, Pacific Grove, CA: Brooks/Cole. 231.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, Fourth edition. Springer. 303.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*, New York and Cambridge: Cambridge University Press. 62–65.

Young, F. W. (1975). An asymmetric Euclidian model for multiprocess asymmetric data. Paper presented at U.S.-Japan Seminar on MDS. San Diego, U.S.A. 79–88.