Journal of the Korean Data & Information Science Society
2009, **20**(2), 301–309

한국데이터정보과학회지

붓스트랩을 이용한 다차원척도법의 효율성 연구[†]

김우종[1] · 강기훈[2]

[12]한국외국어대학교 통계학과

접수 2009년 1월 20일, 수정 2009년 3월 18일, 게재확정 2009년 3월 24일

요 약

다차원척도법은 다변량분석에서 개체들을 대상으로 변수들을 측정한 후에 개체들 사이의 비유사성을 측정하고, 그 값들 혹은 반복하여 측정된 경우에는 그 값들의 평균을 이용하여 개체들을 저차원의 공간상에 도시화시켜 표현하는 분석방법이다. 본 논문에서는 응답자의 답변에 기초하여 비유사성을 측정할 때 이상치 또는 응답자의 답변이 불성실할 경우 발생하는 변이문제와 개체들 간의 거리에 대한 통계적 추론 문제에 붓스트랩 방법을 적용하는 내용을 다루고, 활용가능성을 무료일간지에 대한 유사성 평가 자료를 이용하여 실증적으로 분석하였다.

주요용어: 다변량분석, 비유사성, 유사성, 이상치.

1. 서론

다차원척도법(multidimensional scaling; MDS)은 마케팅 분야에서 소비자의 상표간 유사성 판단 혹은 상표에 대한 선호판단 등에 근거하여 경쟁분석의 중요한 방법인 포지셔닝(positioning)을 수행하는 다변량 분석 기법의 일종이다. 관련된 속성에 대하여 소비자에게 질문을 하는 것이 아니라 대상간의 유사성을 질문하여 그 유사성 자료를 분해하고 평가대상간의 관계를 다차원 공간에 표시하여 준다. 다차원척도법은 집단자료(aggregated)의 평균값을 많이 이용하는데 설문조사로 얻어진 개체들의 비유사성 정도를 평균값을 내어 사용하는 것이다. 이러한 조사에서 소비자의 응답이 불성실할 경우 자료의 분산은 커지게 되고 그 자료의 신뢰성에 문제가 생긴다. 특히, 이상치(outlier)가 존재하는 경우에는 더욱 심각해진다. 다차원척도법의 응용사례에 관한 연구로는 Hwang과 Park (2003), Koh와 Lee, (2004) 등이 있다.

다차원척도법의 하나인 유사성 다차원척도법의 기본원칙은 심리적 개념인 비유사성과 공간개념인 거리가 비슷하다고 한 Richardson (1938)의 주장에 기초하고 있다. 소비자들이 대상들을 심리적으로 인지하고 있는 비유사성을 자료로 하여 공간상의 위치로 전환시켜 그 대상들 간의 거리를 구하여 유사성을 판단하는 것이다. 하지만 그 거리는 그에 대한 변이를 모르는 상태에서는 확증적 자료 분석의 결과로 사용하기에는 미흡하다고 하겠다. 또한, 심리적인 개념인 비유사성이 이러한 공간개념과 잘 맞는지에 대해서도 단정적으로 결론을 내릴 수는 없다.

이러한 측면에서 본 논문은 자료가 분산이 크거나 이상치가 존재할 때 붓스트랩 기법을 활용하여 개체들의 거리와 추정된 거리 사이의 적합도를 측정하는 지표인 표준화잔차제곱합(Standardized Residual

[†] 이 논문은 2008년 한국외국어대학교 학술연구비 지원에 의해 이루어졌음.
[1] (130-791) 서울시 동대문구 이문동, 한국외국어대학교 대학원 통계학과, 석사과정.
[2] 교신저자: (449-791) 경기도 용인시 처인구 모현면, 한국외국어대학교 정보통계학과, 부교수.
E-mail: khkang@hufs.ac.kr

Sum of Squares: STRESS)을 얼마나 안정적으로 낮출 수 있는지 살펴보고, 대상들 간의 거리 설명력을 붓스트랩 신뢰구간을 통해 해석하는데 그 목적이 있다. STRESS값이 클 경우 표현된 최적 위치의 적합성은 그 기준에 의해 신뢰할 수 없는 결과를 나타낸다. 우선, 다차원척도법에 대한 일반적인 내용을 살펴보고, 무료 일간지에 대한 변이가 큰 유사성 평가 자료를 이용하여 대안으로 제시된 붓스트랩 방법을 적용했을 때 다차원척도법의 효율성, 그리고 대상간의 거리를 붓스트랩 신뢰구간을 통해 실증적으로 해석, 활용가능성을 제시하고자 한다.

2. 다차원척도법

2.1. 유사성 다차원척도법

많은 경우 자료분석에서 잘 작성된 하나의 그림이 여러 개의 수치들보다 더욱 가치가 있을 수 있다. 이것은 그림으로 표현된 정보가 수치로 표현된 정보보다 훨씬 이해하기가 쉽다는 것을 의미한다. 하지만, 자료가 사람들이 특정 대상들을 평가한 것이라면, 그 대상에 대해서 사람들이 어떻게 느끼고 있는가를 알기란 쉽지 않을 뿐만 아니라 결과를 그림으로 명료하게 표시한다는 것 또한 어렵다.

다차원척도법은 응답자가 느끼고 있는 다양한 측면의 지각도나 선호도를 좌표상의 그림으로 표현하는 방법으로써 이러한 문제들을 어느 정도 해결해준다. 응답자의 심리적 또는 추상적인 개념들을 저차원의 공간에 표현하여 느낌들에 대한 관계를 보다 쉽게 파악할 수 있도록 평면상에 상대적 거리로 나타내는 것이다. 수학적으로는 2차원공간이나 3차원 공간뿐만 아니라 n차원공간에서도 다차원척도법을 실행할 수 있으나, n이 커짐에 따라 그림으로 나타내는 방법이 힘들어진다.

이렇게 표현한 위치도에서 서로 가까운 곳에 위치한 개체들은 상대적으로 유사함을 의미하게 된다. 결과적으로 얻어진 기하학적 공간을 위치도(positioning map) 혹은 인지도(perception map)라고 하는데 그 공간을 구성하는 차원이 평가기준이 되며 그 차원상의 좌표가 각 대상의 평가수준이 된다. 유사성 다차원척도법에는 개체간의 실제 측정 거리값이나 유클리드 거리를 이용하는 계량형 다차원척도법(metric MDS)과 개체간의 실제 측정 거리값이나 유클리드 거리의 크기 순서를 이용하는 비계량형 다차원척도법(non-metric MDS)이 있는데 본 연구에서는 실제 유클리드 거리를 이용하는 것이기 때문에 계량형 방법에 해당된다.

p개의 변수로부터 관측된 n개의 개체(object)에 대해 얻은 다변량 자료 행렬을

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_n} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \{x_{ir}\}, \quad i = 1, 2, \cdots, n, \quad r = 1, 2, \cdots, p$$

라 했을 때, 유사성 다차원척도법에서는 p차원의 유클리드 공간에서의 n개의 개체간의 비유사성 (dissimilarity) $d_{ij}(i, j = 1, ..., n)$를 리커트 척도(Likert scales)로 측정한 거리를 사용한다. 따라서 다차원척도법은 n개의 개체간의 비유사성을 나타내는 크기가 $n \times n$인 비유사성 행렬(dissimilarity matrix)을 식 (2.1)과 같이 구하고 이 비유사성 행렬 D를 저차원 공간에 기하학적으로 나타낸다. 여기서 비유사성 행렬 D는 대각원소 $d_{ii} = 0$이고 $d_{ij} = d_{ji}$인 대칭행렬이다.

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} = \{d_{ij}\}, \quad i,j = 1, \cdots, n, \tag{2.1}$$

개체 i와 j간의 유클리드 거리(Euclidean distance)는 다음과 같다.

$$d_{ij} = \sqrt{\sum_{r=1}^{p}(x_{ir} - x_{jr})^2}, \quad i,j = 1, \cdots, n.$$

이 식을 개체 i와 j사이의 비유사성으로 정의하고, d_{ij}가 클수록 개체 i와 j가 유사하지 않음을 나타낸다. 유사성 다차원척도법은 다차원척도법의 기원이며, 아직까지도 대표적인 모형으로 가장 많이 활용되고 있다. 유사성 다차원척도법은 평가 대상들간의 유사성 또는 비유사성 자료를 근거로 하여 이들 대상들이 소비자에게 인지되고 있는 상태를 종합적으로 한 차원상의 공간에 위치로 전환하여 보여준다. 이러한 유사성 또는 비유사성을 이용한 자료를 입력하게 되면 다차원척도법은 그림 상에 각 평가대상들에 해당하는 점을 찍어주게 되는데 이것이 바로 각 평가대상들이 지각되고 있는 위치인 것이다. 유사성 다차원척도법의 적용 사례로는 김영찬과 김주영 (2000), 유도재와 김성혁 (2005) 등이 있다.

2.2. 표준화잔차제곱합(STRESS)

다차원척도법은 개체들 사이의 비유사성을 이용하여 공간상에 개체를 표현할 때, 개체들 사이의 비유사성 정도를 최적으로 표현하기 위한 반복과정을 통해서 이루어지게 된다. 표현된 최적 위치의 적합성은 Kruscal과 Wish (1978)의 표준화잔차제곱합(STRESS)을 이용한다. Kruscal의 STRESS는 공간상의 표현이 주어진 비유사성에 어느 정도 적합한가를 측정하는 기준이 되며 다음과 같이 정의된다.

$$\text{STRESS} = \sqrt{\frac{\sum_{i<j}(d_{ij} - f(\delta_{ij}))^2}{\sum_{i<j}(d_{ij})^2}}$$

여기서 δ_{ij}는 소비자가 평가한 대상 i와 j간의 비유사성 정도를 나타내고, d_{ij}는 대상 i와 j간의 거리를 나타낸다. STRESS에서 분자는 공간상의 d_{ij}와 비유사성의 함수 $f(\delta_{ij})$사이의 차이 제곱을 나타내고, 분모는 일반적으로 차원이 증가하면 거리 d_{ij}는 증가하기 때문에 서로 다른 차원의 적합도를 비교하기 위하여 사용되는 표준화 역할을 한다. 만일 두 거리가 동일하면, 즉 $d_{ij} = f(\delta_{ij})$이면 STRESS값은 0이 되고 d_{ij}는 δ_{ij}의 함수에 의해 완벽하게 추정된다는 것을 의미한다. 반대로 STRESS 값이 크다면 실제 거리 d_{ij}와 추정된 거리 $f(\delta_{ij})$의 차이가 크다는 의미가 된다. 최적모형의 적합은 부적합도(badness-of-fit)인 STRESS를 최소로 하기 위한 최적화 알고리즘을 이용하며, STRESS 값이 일정한 수준이하로 될 때 최종적으로 적합된 모형을 제시하게 된다. STRESS 값은 0과 1사이의 값을 취하며, 0에 가까울수록 적합된 모형이 적절하다고 판단한다. 차원수 결정을 위한 Kruscal의 STRESS 판별기준 (Kruscal, 1978)은 표 2.1과 같다.

3. 붓스트랩을 활용한 다차원척도법

3.1. 실험설계

다차원척도법은 대상들간의 유사성을 평가하게 하고 평가자가 대상을 평가하는데 내재하고 있는 평가

표 2.1 KRUSCAL의 STRESS 판별기준

STRESS	0	0.05 이내	0.05∼0.10	0.10∼0.15	0.15 이상
의 미	완벽(perfect)	뛰어남(excellent)	좋음(good)	보통(fair)	나쁨(poor)

기준을 발견해서 각 기준에 따라 평가대상들을 다차원 공간상에 나타내어 주는데 목적이 있다. 이러한 원리에 의하여 본 논문에서 다루고자 하는 무료일간지의 다차원척도분석은 그것이 어떻게 포지셔닝되어 있는지를 평가하고 무료일간지에 관하여 구독자들이 지각하고 있는 모습을 확인하는 과정에 해당된다.

분석 자료는 개인들마다 큰 편차를 보일 가능성이 있는 상표간의 유사성 자료를 분석하기로 하고, 무료일간지 신문 6개를 분석대상으로 ${}_6C_2$개의 짝에 대한 유사성을 응답자에게 평가하도록 하였다. 무료일간지의 독자를 대상으로 하여 리커트 7점 척도를 기준으로 설문조사를 실시하였으며, 조사 대상은 신문은 2008년 9월 기준 현재 배포되고 있는 무료일간지를 연구대상으로 선정하였다. 선정된 신문으로는 포커스, 메트로, AM7, ZOOM, 노컷뉴스, 굿모닝서울 6개이다. 조사방법은 편의표본추출법으로 응답자가 직접 기입하는 자기기입식 방법을 통해 이루어졌다. 자료들의 표본 크기는 50이었으며, 표본의 분포 특성은 표 3.1과 같다. 본 논문에서 사용된 다차원척도법 분석과 붓스트랩 관련해서 소프트웨어는 모두 SAS 9.1 버전을 사용하였다. 조사되어진 ${}_6C_2$개의 짝에 대한 유사성 응답자 평가 자료를 행렬 형태로 전부 입력한 후 그 자료를 surveyselect 프로시져를 통해 복원추출해서 평균을 낸 후 다시 그 값들의 평균을 이용해서 mds 프로시져를 수행하였다. 유사성 자료들을 다시 크기가 50과 10인 표본으로 각각 100회, 500회, 1000회 붓스트랩 기법으로 추출한 뒤, 이들로부터 구한 STRESS값을 상호 비교해 보았으며, 실질적인 내용은 3.2절과 3.3절에 서술되어 있다. SAS를 이용한 다차원척도법의 일반적인 적용에 관해서는 최용석 (1995)을 참고할 수 있다.

표 3.1 표본의 분포 특성

성 별	빈도(%)	직 업	빈도(%)	연 령	빈도(%)
남성	28(56)	대학(원)생	39(78)	10대	7(14)
여성	22(44)	회사원	9(18)	20대	41(82)
		기타	2(4)	30대	2(4)

3.2. 자료분석

표 3.2는 무료일간지 신문의 유사성 행렬 자료로서, 값이 클수록 무료일간지의 유사성의 정도가 상대적으로 커짐을 의미하며, 작을수록 상이함을 나타낸다. 이 자료를 다차원척도법을 통해 이차원 평면상에 나타낸 위치도가 그림 3.1이다. 하지만 표 3.3의 결과에서 보면 C와 D, C와 E, E와 F의 경우 분산이 매우 크게 나왔으므로 이 자료의 안정성은 문제가 있다고 볼 수 있다. 특히 C와 D에 관한 응답에는 이상치가 있는 경우로 보여 그 문제가 더 심각하다. 이처럼 다차원척도법은 그 결과를 대부분의 경우 이차원의 그래프로 표현하기 때문에 이에 따른 외적안정성, 즉 자료수집시 소비자의 응답이 불성실할 경우에 발생할 수 있는 변이를 다루는 문제는 매우 중요하다. 이러한 문제를 다차원척도법에서의 외적안정성 기준인 STRESS값을 통해 살펴보고 붓스트랩을 이용하여 STRESS를 줄일 수 있는지를 확인해 보자.

이 자료를 이용해 SAS로 다차원척도법을 수행한 결과 STRESS값은 0.1286으로 나타났다. 표 2.1에 제시된 Kruscal의 STRESS 판별기준에 의하면 이 자료의 설명력은 좋지 않은 편에 속한다. 부정확한

표 3.2 무료일간지 신문의 유사성행렬

	포커스(A)	메트로(B)	AM7(C)	ZOOM(D)	노컷뉴스(E)	굿모닝서울(F)
포커스	-					
메트로	6.30	-				
AM7	4.60	3.20	-			
ZOOM	3.50	2.72	4.60	-		
노컷뉴스	5.12	5.30	5.56	5.40	-	
굿모닝서울	3.96	4.34	4.06	3.68	3.90	-

표 3.3 무료일간지 신문의 자료수집 결과

	AB	AC	AD	AE	AF	BC	BD	BE	BF	CD	CE	CF	DE	DF	EF
평 균	6.3	4.6	3.5	5.12	3.96	3.2	2.72	5.3	4.34	4.6	5.56	4.06	5.4	3.68	3.9
분 산	0.54	1.71	1.24	1.37	1.79	1.51	1.35	1.97	1.98	5.67	4.09	2.18	1.67	1.57	4.21

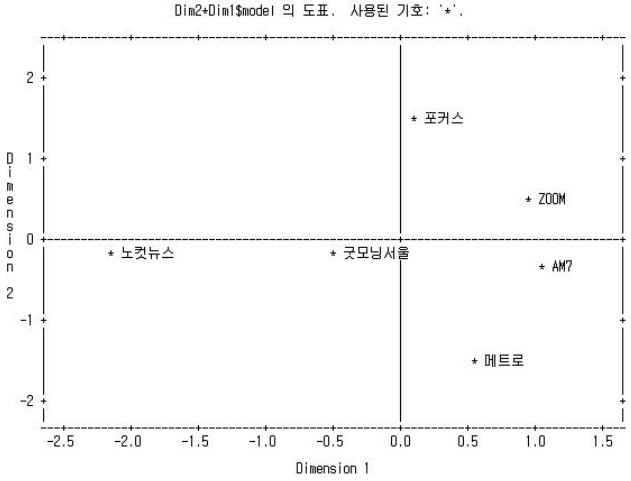


그림 3.1 무료일간지 신문의 위치도

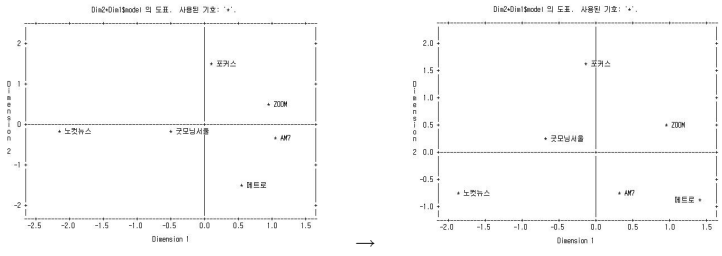
응답으로 인한 변이가 이러한 문제점을 야기했다고 볼 수 있는데, 이를 해결하기 위해 방편으로 붓스트랩 기법을 적용해 보자.

소비자들이 평가한 무료일간지간의 비유사성 평가 자료를 $x = (x_1, x_2, \cdots, x_n)$라고 하고, x로부터 랜덤복원추출한 붓스트랩 표본을 $x^* = (x_1^*, x_2^*, \cdots, x_n^*)$라고 표기하자. 이 자료로부터 독립적으로 각각 $B = 100, 500$ 그리고 1000회 반복하여 붓스트랩 표본 $x^{*^1}, x^{*^2}, \cdots, x^{*^B}$를 뽑는다. 이렇게 얻은 붓스트랩 표본들에 대해 구하고자 하는 추정량인 평균 $\bar{x}_b^*, b = 1, \cdots, B$를 계산한다. B회의 붓스트랩 반복으로 구한 추정량인 평균을 $\sum_{i=1}^{B} \bar{x}_i^*/B$의 식을 이용하여 구하고, 그 값으로 다차원척도법을 실행하여 얻은 STRESS값이 표 3.4와 같다. 표 3.4를 보면 STRESS값은 0.0538에서 0.0555까지 정도로 원자료에서 구한 0.1286에 비해 붓스트랩 방법을 이용한 후 이 자료의 설명력은 뛰어남을 알 수 있다. 또한 붓스

트랩 표본크기를 50과 10으로 추출해 비교해 보았을 때, 붓스트랩 재표본의 크기가 원래 표본의 크기보다 적은 10인 경우도 STRESS값은 여전히 좋게 나타남을 알 수 있다. 물론, 구현 알고리즘에 따라 적합도가 달라질 수 있고 Kruskal 판별기준에 기초한 해석이 절대적인 것이 될 수는 없으나 적어도 붓스트랩을 사용해서 나빠지지 않는다는 사항에는 이견이 없다고 할 수 있다.

표 3.4 붓스트랩 후 STRESS값

반복횟수	붓스트랩 표본크기	STRESS값
100	50	0.0554
	10	0.0551
500	50	0.0554
	10	0.0548
1000	50	0.0555
	10	0.0538



(a) 붓스트랩 전 (b) 붓스트랩 후

그림 3.2 붓스트랩 전, 후의 위치도

그림 3.2는 무료일간지 유사성 자료에 대해서 원자료의 위치도와 붓스트랩 반복수 1000과 붓스트랩 표본크기를 50으로 하여 얻은 자료의 위치도를 비교한 것이다. 원자료의 위치도 같은 경우, STRESS값이 높은 자료를 사용했으므로 자료의 안정성에 문제가 있고, 그에 따른 위치도의 신뢰도도 문제가 있다고 볼 수 있다. 표 3.3에서의 분산이 높은 경우는 C와 D, C와 E, E와 F였다. 즉 AM7(C)과 노컷뉴스(E)의 자료 설명력이 낮은 것을 알 수 있고, 그림 3.2의 붓스트랩 전,후의 위치도를 비교해보면 다른 대상들에 비해 AM7, 노컷뉴스와 메트로의 위치가 많이 바뀐 것을 볼 수 있다.

붓스트랩 후의 위치도를 기준으로 분석결과를 해석해보면 차원1(Dimension 1)은 우측을 기준으로 메트로, ZOOM, AM7, 포커스, 굿모닝서울, 노컷뉴스 순으로 나타났는데, 이는 신문에 대한 선호도라고 판단되어진다. 우측에 위치한 신문일수록 더 선호하는 것을 의미하고, 메트로와 ZOOM이 제일 우측에 포지셔닝 되어 있으므로 선호도가 가장 높은 것으로 나타났다. 이 결과는 실제로 2007년 9월호의 대학내일의 선호도조사 결과와 일치한다. 즉, 메트로가 대학생 선호도 1위로 조사됐고, ZOOM 같은 경우는 선호도증감률에서 1위로 조사됐다. 이에 반해 노컷뉴스는 사설, 칼럼이 있는 저널리즘 신문으로써 대학생들에게는 선호도를 얻지 못하는 것으로 나타났다. 원자료의 위치도에서는 AM7이 가장 선호되는 것으로 나타났지만 붓스트랩 후의 위치도에서는 올바르게 적합도를 높여주기 위해 좌측으로 이동하였음을 알 수 있다. 반면 메트로의 경우에는 우측으로 이동하였다.

차원2(Dimension 2)는 신문의 이미지를 결정하는 타이틀의 색을 결정하는 요인이라고 판단된다. 타이틀의 색이 빨간 계열(포커스: 빨강, ZOOM: 주황, 굿모닝 서울: 주황)일수록 위쪽으로 포지셔닝 되어있고, 그렇지 않을수록(노컷뉴스: 파랑, 메트로: 초록, AM7: 혼합) 아래쪽으로 포지셔닝 되어져있

다. 그림 3.2에서 붓스트랩 전, 후의 위치도를 비교해 보면 노컷뉴스와 AM7이 아래쪽으로 큰 이동을 하였는데, 이는 차원2의 속성에 따라 노컷뉴스와 AM7색이 빨간 계열이 아니기 때문에 아래쪽으로 이동된 것이라 여겨진다.

이러한 결과는 붓스트랩 기법을 사용해 STRESS값을 낮춤으로서 AM7(C)과 노컷뉴스(E)의 자료 설명력을 높이고, 이를 통해 붓스트랩 기법에 의해 위치도를 도시화 하는 것의 유용성을 보여주는 것이다.

3.3. 붓스트랩 거리 신뢰구간

다차원척도법의 결과를 차원축소의 의미에서 이차원으로 표현할 경우 각 대상들간의 거리에 대한 통계적 추론은 매우 의미가 있다. 다차원척도법에 있어서 어느 대상들끼리 더 가까운지 알아보는 것은 다차원척도법의 해석에 있어서 중요한 문제이기 때문이다.

이차원상에 나타내어진 대상들 간의 거리에 대한 추론을 하기 위하여 다차원 자료를 이차원상에 근사되어진 두 대상 $i(a_i, b_i)$와 $j(a_j, b_j)$의 거리를

$$d_{ij} = \sqrt{(a_i - a_j)^2 + (b_i - b_j)^2}$$

라고 하자. 이 d_{ij}는 두 대상들간의 실제거리는 아니지만 이차원상에 근사된 거리로서 대상들간의 거리에 대한 추론에 이용될 수 있다. 이 추론에 대한 신뢰도를 높이기 위해서는 좌표점 간의 거리에 대한 변이를 알아야 한다. 구간추정을 통하면 이러한 대상들 간의 거리가 빈번히 포함되어 있을 것으로 생각되는 범위를 추정할 수 있다. 그러므로 본 논문에서는 d_{ij}의 거리를 붓스트랩 신뢰구간을 통해 구하고자 한다. 임의의 두 대상 좌표점 사이의 거리 d_{ij}에 대한 붓스트랩 표본의 거리를

$$d^*_{ij} = \sqrt{(a^*_i - a^*_j)^2 + (b^*_i - b^*_j)^2}$$

로 두면 이차원상에 근사된 대상들 간의 거리에 대한 추론은 붓스트랩 신뢰구간을 이용하여 가능해진다. 붓스트랩 신뢰구간 계산에 필요한 모의실험의 단계는 다음과 같다.

- **1단계:** 소비자들이 평가한 무료일간지의 비유사성 평가 자료인 $x = (x_1, x_2, \cdots, x_n)$로 부터 랜덤복원 추출한 붓스트랩 표본 $x^* = (x^*_1, x^*_2, \cdots, x^*_n)$을 $B = 100$회 반복하여 추출한다.
- **2단계 :** 각 ${}_nC_2$의 쌍에 대한 붓스트랩 표본 $x^{*^1}, x^{*^2}, \cdots, x^{*^B}$로부터 추정량인 평균 $\bar{x}^*_b$, $b = 1, \cdots, B$를 계산한다.
- **3단계:** 2단계의 평균 $\bar{x}^*_b$를 이용하여 다차원척도법을 수행하고 붓스트랩 거리 d^*_{ij}를 $B = 100$회 계산한다.
- **4단계:** 상하 95% 백분위를 구하여 95% 붓스트랩 신뢰구간을 구한다.

앞에서 다루었던 무료일간지의 유사성 자료에 대하여 좌표점간의 거리에 대한 신뢰구간을 붓스트랩 방법으로 구한 결과는 표 3.5와 같다. 원자료의 경우 STRESS값이 높으므로 원자료를 이용한 거리는 신뢰할만한 결과가 아니다. 붓스트랩 후의 표본을 이용한 거리가 신뢰할만 하지만 붓스트랩 후 거리도 그에 대한 변이를 모르는 상태에서는 확증적 결과로 사용하기에 미흡하다. 붓스트랩 신뢰구간 방법을 이용하면 이러한 문제를 해결할 수 있다. 여기서 대상들 간의 참 거리가 1.0 이내라고 간주할 수 있는 좌표 점들은 B와C, B와D, C와D, E와F 이며, 이는 그림 3.2의 붓스트랩 후 위치도의 결과와 일치되는 모습이다.

원자료에서는 C와D의 거리가 가장 가깝고, C와E의 거리가 가장 멀게 나타났지만, 붓스트랩 신뢰구간에서는 B와C의 거리가 가장 가깝고, B와E의 거리가 가장 멀게 나타난 것을 볼 수 있다. 이는 자료의 안정성의 문제가 있을 때, 즉 STRESS값이 높게 나타났을 때 원자료와 붓스트랩 후의 자료의 결과는 아

표 3.5 각 대상들의 거리에 대한 붓스트랩 95% 신뢰구간

	원자료거리	붓스트랩거리	하한값	상한값
AB	2.942	2.873	2.791	3.085
AC	2.081	3.013	2.432	3.034
AD	1.253	1.755	1.481	1.986
AE	2.817	2.899	2.711	3.112
AF	1.693	1.764	1.487	2.018
BC	1.184	1.093	0.733	1.427
BD	2.033	1.182	0.899	1.545
BE	2.965	3.011	2.875	3.355
BF	1.701	2.209	2.096	2.642
CD	0.933	1.761	0.898	1.908
CE	3.223	2.103	1.652	2.543
CF	1.599	1.701	1.331	1.901
DE	3.206	2.908	2.752	3.058
DF	1.612	1.761	1.509	2.055
EF	1.646	1.239	0.912	1.581

주 다르게 나타날 수 있음을 보여 주는 것이다. 즉, 붓스트랩 기법을 통해 STRESS값을 낮추면서 그에 대한 자료의 설명력을 높이고, 붓스트랩 신뢰구간을 통해 그에 대한 신뢰도를 높일 수 있다.

4. 결론

다차원척도법에 사용되는 유사성행렬 자료는 집단자료의 평균값을 이용한다. 하지만 집단자료의 평균값으로 다차원척도법을 수행하는 경우 자료에 이상치가 있다면 그 결과의 안정성에 문제를 줄 수 있다. 붓스트랩 기법은 이러한 자료들의 안정성을 검토할 수 있는 방법이다. 자료분석 결과 분산이 큰 자료를 가지고 다차원척도법을 실행하는 경우 적합성 지표인 STRESS 값을 증대시키는 결과를 초래하고 있는데 이것은 안정성에 심각한 문제를 주고 있다는 것을 의미한다. 특히 이상치가 있는 경우는 그 정도가 더 심각하다.

본 논문에서는 이러한 문제점에 대해 붓스트랩 기법의 활용가능성을 모의실험을 통해 살펴보았다. 제한된 실험이기는 하지만 그 결과 STRESS 값이 높을 경우 붓스트랩 기법으로 그 문제점을 해결할 수 있다는 것을 보였고, 각 대상들 간의 거리에 대한 통계적 추론에 대해 붓스트랩 신뢰구간을 통해 신뢰도를 높일 수 있었다. 따라서 다차원척도법에서 자료에 근거한 붓스트랩 기법의 활용이 유익할 것으로 기대되며 본 논문에서 다루지는 않았지만 이에 대한 이론적인 사항도 연구할 필요가 있을 것으로 생각된다.

참고문헌

김영찬, 김주영 (2000). 다차원척도법의 활용방안 및 발전방향. <소비자학연구>, **11**, 199-227.

유도재, 김성혁 (2005). 호텔 포지셔닝 분석에 있어 다차원척도법의 적용. <관광연구저널>, **19**, 99-111.

최용석 (1995). <SAS 다차원척도법>, 자유아카데미.

Hwang, S. Y. and Park, S. K. (2003). Scaling MDS for preference data using target configuration. *Journal of the Korean Data & Information Science Society*, **14**, 237-245.

Koh, B. S. and Lee, G. E. (2004). Web log analysis system using SAS/AF. *Journal of the Korean Data & Information Science Society*, **15**, 317-329.

Kruskal, J. B. and Wish, M. (1978). Multidimensional scaling. *Sage University Paper Series on Quantitative Applications in the Social Science.*

Richardson, M. W. (1938). Multidimensional psychophysics. *Psychological Bulletin*, **35**, 659-660.

Journal of the Korean Data & Information Science Society
2009, **20**(2), 301–309

한국데이터정보과학회지

A study on the efficiency of multidimensional scaling using bootstrap method[†]

Woojong Kim[1] · Kee-Hoon Kang[2]

[12]Department of Statistics, Hankuk University of Foreign Studies

Received 20 January 2009, revised 18 March 2009, accepted 24 March 2009

Abstract

Multidimensional scaling(MDS) is a statistical multivariate analysis technique that is often used in information visualization for exploring similarities or dissimilarities in data. In order to analyse and visualize data, MDS measures the dissimilarities between objects and uses them or their mean if they are repeatedly measured. When there exist outliers or when the variation of data is too large, we can hardly get reliable results on the research using MDS. In this paper, we consider the MDS based on bootstrap method when the variation of data is large. Standardized residual sum of squares is considered as measuring goodness-of-fit of the model. A real data analysis is included to examine our approach.

Keywords: Dissimilarity, multivariate analysis, outlier, similarity.

[†] This research was supported by the research fund of Hankuk University of Foreign Studies, 2008.
[1] Graduate Student, Department of Statistics, Hankuk University of Foreign Studies, Seoul 130-791, Korea.
[2] Corresponding author: Associate Professor, Department of Statistics, Hankuk University of Foreign Studies, Mohyeon, Cheoin-goo, Yongin 449-791, Korea. E-mail: khkang@hufs.ac.kr