응용통계연구 제 **14**권 2호, *2001*년, *pp. 401–413*

대응분석에 있어서 붓스트랩 방법의 활용에 대한 고찰

강창완 [1] 김대학 [2] 전명식 [3]

요 약

이차원 분할표자료에 대해서 행과 열의 관계를 저차원상에 시각적으로 표현하는 탐색적대응분석에 대하여 붓스트랩방법의 사용가능성을 살펴보았다. 기존의 탐색적 면만이 강조되어 왔던 대응분석에서 좌표점의 변이와 좌표점간의 거리에 대한 통계적 추론을 붓스트랩방법으로 해결할 수 있음을 보이고 또한 좌표축의 설명력에 대하여 붓스트랩신뢰구간의 포함확률의 일치성을 모의실험을 통해 제시하였다.

주요용어: 분할표, 대응분석, 붓스트랩, 통계적 추론, 포함확률.

1. 서 론

대응분석(correspondence analysis)은 분할표 자료의 행과 열의 관계를 표현하는 다차원 척도법의 일종이다. 이는 행과 열을 통상적으로 이차원 그래프상의 점으로 나타내고 그들 사이의 관계를 알아보는데 사용되는 일종의 탐색적 통계기법으로 확증적 자료분석의 결과로는 미흡한 점이 많다. 본 논문에서는 분할표 자료의 분석에 사용되는 대응분석에 붓스트랩방법을 활용하여 대응분석의 변이와 저차원상에 도시된 표본 좌표점간의 거리, 그리고 좌표축의 설명력을 확증적으로 해석하고 그 활용가능성을 모의실험을 통하여 제시하고자 한다.

표 1.1과 같은 이차원 분할표의 형태로 주어진 크기가 $I \times J$ 인 자료행렬 $N = \{n_{ij}\}$ 을 고려하자. 여기서 자료는 다항추출(multinomial sampling model)모형을 가정한다. 즉, 두개의 변인에 대하여 각각 I개와 J개의 범주가 있으며 관찰개체는 $I \times J$개의 칸(cell)들 중의 하나에 속한다. 이 때 n_{ij} 는 i번째 행의 j번째 열에 해당하는 관찰개체의 빈도를 의미하며 편의상 $n_{i.}$ 은 i번째 행의 합 그리고 $n_{.j}$는 j번째 열의 합으로 나타내자. 또 전체 관찰갯수는 $n_{..} = \sum\sum n_{ij}$ 로 표기한다. 이에 대한 대응행렬을

$$P = \frac{1}{n_{..}} N = \{p_{ij}\}$$

로 표기하고 나아가 일반성을 잃지않고 편의상 $I \geq J$를 가정하자. 그러면, 대응행렬 P의 행과 열의 합은 각각

$$\mathbf{r} = P\mathbf{1}_J, \quad \mathbf{c} = P^T\mathbf{1}_I$$

1) (614-714) 부산시 부산진구 가야동 산24, 동의대학교 수학 · 컴퓨터통계학부, 조교수
E-mail: cwkang@hyomin.dongeui.ac.kr
2) (712-702) 경북 경산시 하양읍 금락리 330, 대구가톨릭대학교 정보통계학과, 교수
E-mail: dhkim@cuth.cataegu.ac.kr
3) (136-701) 서울시 성북구 안암동 5, 고려대학교 통계학과, 교수
E-mail: jhun@korea.ac.kr

표 1.1: 이차원 분할표

	열 1	열 2	$\cdots$	열 J
행 1	n_{11}	n_{12}	$\cdots$	n_{1J}
행 2	n_{21}	n_{22}	$\cdots$	n_{2J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
행 I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}

가 된다. 여기서 $\mathbf{1}_J$ 와 $\mathbf{1}_I$는 각각 1을 원소로 갖는 각각 $J \times 1$ 행렬, $I \times 1$ 행렬이다. 또한 $\mathbf{r}$과 $\mathbf{c}$의 원소를 대각원소로 갖는 대각행렬을

$$D_r = \mathrm{diag}(\mathbf{r}), \quad D_c = \mathrm{diag}(\mathbf{c})$$

로 표기하자. 그러면, 행 프로화일 행렬과 열 프로화일 행렬은 각각

$$R = D_r^{-1}P = (\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_I)^T, \quad C = D_c^{-1}P^T = (\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_J)^T$$

이 되는데 이 프로화일들은 관찰빈도를 해당하는 행 또는 열의 합으로 나눈 결과와 같다. 즉,

$$\mathbf{r}_i^T = (n_{i1}/n_{i\cdot}, \cdots, n_{iJ}/n_{i\cdot}), \quad \mathbf{c}_j^T = (n_{1j}/n_{\cdot j}, \cdots, n_{Ij}/n_{\cdot j})$$

이 된다.

다음으로 조정된 대응행렬 $P - \mathbf{r}\mathbf{c}^T$에 대한 특이치분해(singular value decomposition)

$$P - \mathbf{r}\mathbf{c}^T = AD_\mu B^T$$

를 구한다. 이때 $A^T D_r^{-1} A = B^T D_c^{-1} B = I$, $D_\mu = \mathrm{diag}(\mu)$, $\mu = (\mu_1, \cdots, \mu_k)$, $\mu_1 \geq \cdots \geq \mu_k > 0$ 이다. 그러면, 행 프로화일벡터와 열 프로화일벡터에 대한 주좌표행렬은

$$F = D_r^{-1}AD_\mu, \quad G = D_c^{-1}BD_\mu$$

로 계산된다. 또한, 이들은 각각

$$F = D_r^{-1}PGD_\mu^{-1}, \quad G = D_c^{-1}P^T F D_\mu^{-1}$$

으로 표현될 수도 있다. 그러나, 이렇게 구한 표본좌표점과 각 축에 해당하는 고유값은 확증적 자료분석의 결과로는 미흡한 점이 많다. 이러한 결과는 위의 분석과정이 전적으로 행 프로화일과 열 프로화일에만 의존하며 전체 관찰개체의 수 $n_{\cdot\cdot}$ 과는 무관하다는데 있다. 즉, 전체 관찰갯수의 크기인 $n_{\cdot\cdot}$에 관계없이 같은 행 프로화일(또는 열 프로화일)을 가지면 같은 주좌표들을 제공하게된다. 그러나, 분할표를 구하는데 사용한 표본의 모분포를 가정한다면 구한 주좌표들에 대한 통계적 추론을 고려하는 것은 자연스러운 일이다. 여기서는 이차평면에 하나의 점으로 나타난 행 주좌표(또는 열 주좌표)들의 변이와 좌표들 간의 거리분포, 그리고 좌표축의 설명력에 대하여 붓스트랩방법의 활용을 시도하고자 한다

2. 붓스트랩 방법

Efron(1979)에 의해 제안된 붓스트랩방법은 기저분포에 대한 모수적 가정없이도 표본으로부터 재표본을 취하는 방법에 근거하여 표본분포의 추정에 활용되어왔다. 한편, 이러한 다항분포에 대한 붓스트랩방법의 점근적 이론은 Efron(1982), Woodroofe & Jhun(1988)등에 의해 규명되었으며, 대응분석에서의 붓스트랩 활용가능성은 Greenacre(1984)에 의해 제안된 바 있다. 또한 Reiczigel(1996)은 분할표에서의 적합도 검정문제에 대하여 붓스트랩검정을 다루었고, Balbi(1992)는 비대칭대응분석에서의 안정성문제에 대해 붓스트랩 활용을 발표한 바 있다. 이제 대응분석에서 붓스트랩방법의 활용에 대하여 살펴보기로 하자.

$I \times J$ 개의 칸에 확률질량 p_{ij} 를 가진 다항분포로부터 크기가 $n_{..}$ 인 붓스트랩표본을 구하고 그로부터 $D_{r^*}^{-1}P^*$ 과 $D_{c^*}^{-1}P^{*T}$을 계산한다. 이에 근거한 행 프로화일벡터와 열 프로화일벡터에 대한 붓스트랩 주좌표행렬은

$$F^* = D_{r^*}^{-1}P^*GD_{\mu}^{-1}, \quad G^* = D_{c^*}^{-1}P^{*T}FD_{\mu}^{-1}$$

로 계산된다. 한편, 붓스트랩 주좌표행렬을 구함에 있어 $P^* - r^*c^{*T}$에 대한 특이치분해를 통해 A^*와 $D_{\mu}^* = \mathrm{diag}(\mu^*)$, $\mu_1^* \geq \cdots \geq \mu_k^* > 0$ 를 구하고 $F^* = D_{r^*}^{-1}P^*GD_{\mu}^{-1}$ 과 $G^* = D_{c^*}^{-1}P^{*T}FD_{\mu}^{-1}$ 를 계산하면 식별의 문제가 생길 수 있다. 왜냐하면, 주좌표의 계산은 μ_i^* 의 값에만 의존하며 그 방향은 고려하지 않기 때문이다. 이러한 붓스트랩방법의 근사해를 구하기 위한 알고리즘은 다음과 같다.

단계 1 크기가 $I \times J$인 주어진 분할표 $N = \{n_{ij}\}$에 근거하여 다항분포 $(n_{..}, n_{i1}/n_{..}, \cdots, n_{ij}/n_{..}, \cdots, n_{IJ}/n_{..})$ 를 만든다.

단계 2 단계 1에서 만들어진 다항분포로부터 구한 붓스트랩표본 $N^* = \{n_{ij}^*\}$ 에 대한 행 프로화일벡터 ${\mathbf{r}_i^*}^T = (n_{i1}^*/n_{i\cdot}^*, \cdots, n_{iJ}^*/n_{i\cdot}^*)$ 과 열프로화일 벡터 ${\mathbf{c}_j^*}^T = (n_{1j}^*/n_{\cdot j}^*, \cdots, n_{Ij}^*/n_{\cdot j}^*)$ 를 계산 하고 나아가 F^* 과 G^*를 계산한다.

단계 3 단계 2를 B회 독립반복시행하여 붓스트랩방법의 근사해를 구한다.

한편 위 알고리즘은 추출모형을 다항분포로 가정한 경우이며 만일 추출모형이 적다항분포인 경우 즉, 자료가 독립적으로 I 개의 다항분포로 나온 표본을 가정한 경우의 붓스트랩방법은 Balbi(1992)의 알고리즘을 이용할 수 있다.

2.1. 이차평면에서의 플롯에 근거한 외적안정성

대응분석은 그 결과를 대부분의 경우 이차원 그래프로 표현하는 바 이에 대한 외적 안정성, 즉 같은 모집단으로부터 다른 표본을 구한 경우에 발생할 수 있는 변이에 관한 문제는 매우 중요하다. 이러한 변이의 측정은 기존의 탐색적 대응분석에서는 쉽지 않으나 2절에서 설명한 붓스트랩방법을 이용하여 이차평면에 하나의 점으로 나타난 행 프로화일(또는 열 프로화일)의 변이의 측정을 구하고자 한다. 표2.1는 Greenacre(1984)의 자료로서 지위에 따른 개인의 흡연습관을 나타낸 이차원분할표 자료이다.

표 2.1: Greenacre 자료

지위	흡연습관			
	None	Light	Medium	Heavy
Senior manager	4	2	3	2
Junior manager	4	3	7	4
Senior employes	25	10	12	4
Junior employes	18	24	33	13
Secretaries	10	6	7	2

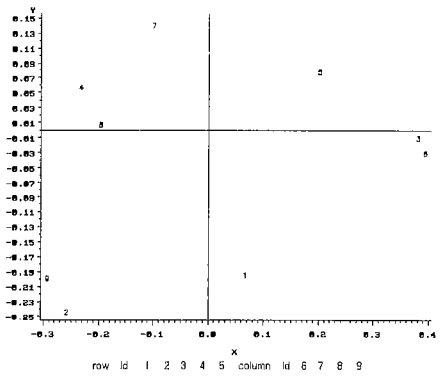

그림 2.1: 행과열 플롯

표2.1의 자료를 단순대응분석을 하여 행과 열을 이차원 평면상에 동시에 플롯한 것이 그림2.1이다. 이러한 행과 열의 대응관계를 탐색적으로 분석하는 대응분석은 전체관찰에서 단순히 상대빈도만을 고려한 것으로서 전체 관찰갯수인 $n_{..}$ 에 관계없이 같은 행 프로화일(또는 열 프로화일)을 가지면 같은 모양의 그림을 제공하게 된다. 즉, 표2.1의 자료를 10배한 자료에 대해서 대응분석을 하여도 같은 이차원 플롯을 가지게된다. 이러한 문제점은 앞에서 언급한 붓스트랩방법을 이용하여 어느정도 해결할 수 있다. 그림2.2은 표2.1의 자료에 대해서 붓스트랩 반복횟수를 $B = 500$ 회로 하여 각 붓스트랩표본에 대한 행과 열을 이차원 평면상에 도시한 그림이다.

한편 표2.1자료에 10배한 이차원 분할표 자료를 붓스트랩방법을 이용하여 행과 열을 각 붓스트랩표본에 대해 플롯한 그림이 그림2.3이다. 그림의 비교를 위해 가로 세로축의 축척은 동일하게 택했다. 그림에서 알 수 있듯이 전체 관찰 갯수가 증가함에 따라 행 프로화일(혹은 열 프로화일)의 변이가 감소하는 것을 알 수 있다. 한편, 다변량자료에서 자료의 변이를 표현하는 방법은 일반적으로 두가지 방법이 널리 알려져 있는데 그 중 하나가 다변량자료로 부터 구한 공분산행렬의 대각원소의 합(trace), 즉 각 변수의 분산들의 합으로 표현하는 방법이고 다른 하나는 일반화 분산이라 불리는 공분산행렬의 행렬식(determinant)으로 표현하는 방법이다. 본 논문에서는 두 가지 방법 모두를 사용하여 프로화일들의 변이를

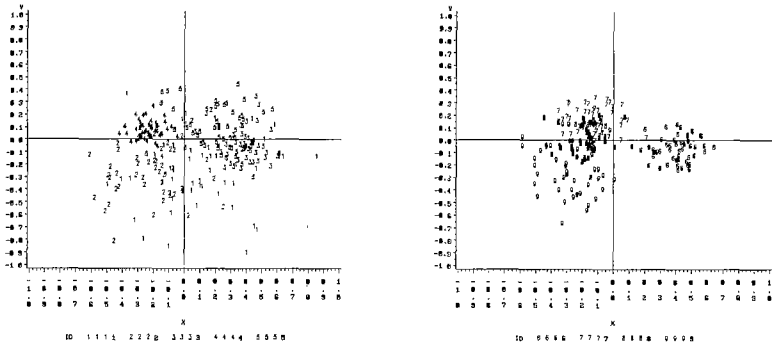

그림 2.2: 행(좌)과 열(우)에 대한 붓스트랩 플롯

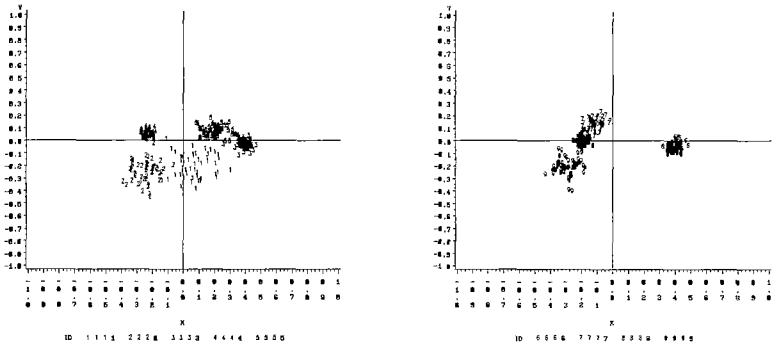

그림 2.3: 10배자료의 행(좌)과 열(우)에 대한 붓스트랩 플롯

살펴보았다. 표2.2은 이러한 관찰갯수의 증가에 따른 행과 열의 각각의 붓스트랩 변이를 공분산행렬의 트레이스(trace)와 일반화분산으로 표현한 것이다.

표2.2 에서 보면 행 프로화일(혹은 열 프로화일)의 이차원 변이(trace 기준)가 모든 행과 열에 걸쳐 전체 관찰갯수가 k배 증가할 수록 변이는 $1/k$씩 감소하는 것을 알 수 있다. 또한 이차원 변이를 일반화분산(행렬식 기준)으로 측정한 경우는 전체 관찰갯수가 k배 증가할 수록 $1/k^2$씩 감소하고 있다. 이와 같은 결과는 통상적인 추정방법에서와 같으며 붓스트랩 방법에 의한 프로화일 변이의 측정의 유용성을 보여주고 있다.

3. 표본좌표점 거리의 표본분포

대응분석의 결과를 차원축소의 의미에서 이차원으로 표현할 경우 각 좌표점간(행간 혹은 열간)에 대한 통계적추론은 매우 의미가 있다. 즉, 탐색적 자료분석의 입장인 대응분석에 있어서 어느 행(혹은 열)좌표점들끼리 더 가까운지 알아보는 것은 대응분석의 해석에 있어서 중요한 문제일 수 있다. 왜냐하면 유사한 행끼리는 가까운 거리를 나타낸다는 점에서 유사행끼리 결합하는 문제라든지 혹은 어느 행 좌표점을 기준으로 했을 때 거리가 비슷

표 2.2: 붓스트랩을 이용한 행과 열의 변이
(단위 Trace : $\times 10^{-2}$, Det. : $\times 10^{-4}$)

	greenacre 자료		10 배자료		50 배자료		100 배자료	
	Trace	Det.	Trace	Det.	Trace	Det.	Trace	Det.
행1	20.93	105.68	1.84	0.83	0.37	0.03427	0.197	0.00956
행2	11.36	30.81	1.16	0.29	0.22	0.01108	0.103	0.00251
행3	2.60	1.65	0.27	0.02	0.06	0.00079	0.027	0.00018
행4	1.18	0.34	0.11	0.01	0.13	0.00016	0.013	0.00004
행5	6.39	9.96	0.64	0.10	0.13	0.00420	0.067	0.00108
열1	2.07	1.07	0.20	0.01	0.04	0.00048	0.024	0.00014
열2	3.19	2.52	0.31	0.02	0.06	0.00098	0.029	0.00020
열3	2.17	1.16	0.21	0.01	0.04	0.00043	0.020	0.00010
열4	7.42	12.78	0.82	0.15	0.15	0.00545	0.077	0.00139

하게 나타난 두 개의 다른 행 좌표점이 있을 때 어떤 행 좌표가 더 통계적의미에서 가까운지를 알아보는 것은 대응분석의 해석에 좀 더 심도있는 결과를 제공할 것이기 때문이다. 이제 앞 절에서 언급했던 주좌표행렬 F의 i번째 행을 f_i라 표기하고

$$f_i^T = (\mathbf{r}_i{}^T G) D_\mu^{-1}$$

라 정의하자. 단 $\mathbf{r}_i{}^T$는 행 프로파일행렬 R의 i번째 행을 의미한다. 마찬가지로 주좌표행렬 G에 대해서도 동일한 방식으로 적용할 수 있다. (주좌표행렬 G의 j번째 행은 g_j로 표기하기로 한다.) 여기서 두 개의 행 프로파일 $\mathbf{r}_i$와 $\mathbf{r}_k$ 사이의 거리(일반적인 χ^2거리)는

$$d^2(\mathbf{r}_i, \mathbf{r}_k) = \sum_{j=1}^{J} (n_{ij}/n_{i\cdot} - n_{kj}/n_{k\cdot})^2 / n_{\cdot j}$$

이고, 두 개의 열 프로화일 c_j 와 c_l 사이의 거리는

$$d^2(\mathbf{c}_j, \mathbf{c}_l) = \sum_{i=1}^{I} (n_{ij}/n_{\cdot j} - n_{il}/n_{\cdot l})^2 / n_{i\cdot}$$

이다. 이 경우 행간 혹은 열간의 거리들은 다음과 같이 주좌표행렬 F 와 G를 이용하여 저차원상에 근사시킬수 있다.

$$d^2(\mathbf{r}_i, \mathbf{r}_k) \approx (f_i - f_k)'(f_i - f_k)$$

$$d^2(\mathbf{c}_j, \mathbf{c}_l) \approx (g_j - g_l)'(g_j - g_l)$$

2차원(저차원)상에 나타난 행(혹은 열) 표본좌표점의 거리에 대한 추론을 하기 위하여 2차원상에 근사되어진 임의의 두 행 좌표점 $i(a_i, b_i)$ 와 $k(a_k, b_k)$의 거리를

$$d_{ik} = \sqrt{(a_i - a_k)^2 + (b_i - b_k)^2}$$

로 두자. 이 d_{ik}는 두 행사이의 실제거리는 아니지만 2차원상에 근사된 거리로서 표본좌표점의 거리에 대한 추론에 이용될 수 있다. 왜냐하면 이렇게 구한 표본좌표 점간의 거리는 그에 대한 변이를 모르는 상태에서는 확증적 자료분석의 결과로는 미흡하기 때문이다. 이제 d_{ik}의 표본분포를 붓스트랩 방법으로 구하기로 한다. 여기서는 앞에서 언급된 $P^* - \mathbf{r}^*\mathbf{c}^{*T}$에 대한 특이치분해를 통해 구한 붓스트랩 주좌표 행렬로부터 붓스트랩 거리의 분포를 유도하여 통계적 추론을 시도하여 보자. 이제 임의의 두 표본 행 좌표점 사이의 거리 d_{ik}에 대한 붓스트랩 거리를

$$d_{ik}^* = \sqrt{(a_i^* - a_k^*)^2 + (b_i^* - b_k^*)^2}$$

로 두면 2차원상에 근사된 표본좌표점 사이의 거리에 대한 추론은 붓스트랩 거리분포를 이용하여 가능해 진다. 여기서 $(a_i^*, b_i^*), (a_k^*, b_k^*)$는 붓스트랩 주좌표행렬 $F^* = D_{r^*}^{-1} P^* G D_\mu^{-1}$ 의 i번째 와 k번째 행 f_i^*,k_i^* 들을 2차원상에 근사시킨 행 좌표점들을 의미한다.

아래의 모의실험을 통하여 포함확률의 관점에서 이러한 붓스트랩 신뢰구간의 타당성을 확인할 수 있다. 여기서 고려된 가상 모집단의 분포 $p_{ij}, i = 1, 2, \cdots, r, j = 1, 2, \cdots, c$ 는 표3.1과 같다. 이 모집단들에서의 행간 그리고 열간의 근사거리는 표3.2, 표3.3와 같다. 이제 행간 열간의 거리에 대한 붓스트랩 신뢰구간 계산에 필요한 모의실험의 단계는 다음과 같다.

(단계1) 모집단으로부터 $(r \times c)$개의 항에 확률질량 $\{p_{ij}\}$ 를 가진 다항분포로부터 크기가 $n_{..}$인 $(r \times c)$ 분할표 표본을 생성한다.

(단계2) 단계 1의 $(r \times c)$ 분할표 표본으로부터 행간 혹은 열간 거리 d_{ik}을 계산한다. 한편 분할표 표본에서 $(r \times c)$개의 항에 확률질량 $\{\hat{p_{ij}}\}$를 추정하고 다시 이로부터 다항분포로부터 크기가 $n_{..}$인 $(r \times c)$붓스트랩 분할표 표본을 생성한다.

(단계3) 단계 2의 $(r \times c)$ 붓스트랩 분할표 표본으로부터 붓스트랩거리 d_{ik}^*를 $B = 500$회 계산하고 백분위(percentile)방법에 의해 붓스트랩 신뢰구간을 구하고 참값의 포함여부를 확인한다.

(단계4) 단계 1,2,3 을 총 1000회 독립반복 시행하여 모집단의 행간 혹은 열간 거리에 대한 붓스트랩 신뢰구간의 포함확률을 계산한다.

표3.4와 표3.5을 보면 표본좌표점간의 거리의 신뢰도가 90%, 95%인 붓스트랩 신뢰구간을 이용한 포함확률이 명목확률에 거의 근사하고 있는 것을 볼 수 있으며 이로부터 붓스트랩 거리분포의 타당성을 확인할 수 있다.

이제 앞에서 다루었던 Greenacre자료에 대하여 좌표점간의 거리에 대한 신뢰구간을 붓스트랩 방법으로 구한 결과는 표3.6과 같다. 여기서 좌표점간의 참 거리가 0.1 이내라고 간주할 수 있는 좌표점들은 행d_{12}, 행d_{13}, 행d_{14}, 행d_{15}, 행d_{24} , 행d_{35}, 열d_{78}, 열d_{79}, 열d_{89} 이며 이는 2.1절의 그림2.2에 보여지는 Greenacre 자료에 대한 붓스트랩플롯 결과와 일치되는 모습이다. 즉, 행 좌표에서는 행2 와 행3, 행2 와 행5 그리고 행3 과 행4 좌표점들이 구분되는 모습을 볼 수 있으며, 열 좌표에서는 열7, 열8, 열9가 열 6과는 확연하게 구분되는 모습을 볼 수 있다.

표 3.1: 가상 모집단 다항분포의 모비율

모집단 I : 3×3			모집단 II : 4×4의 경우			
0.20	0.05	0.05	0.0235	0.0588	0.1412	0.0706
0.10	0.20	0.10	0.0235	0.0118	0.0353	0.1941
0.05	0.10	0.15	0.1471	0.0176	0.0176	0.0412
-	-	-	0.1059	0.0588	0.0412	0.0188

표 3.2: 모집단 I의 행(열) 간의 거리

	행 1	행 2	행 3		열 1	열 2	열 3
행 1	0	0.951	1.079	열 1	0	0.9404	0.9884
행 2	0.915	0	0.555	열 2	0.9404	0	0.5477
행 3	1.079	0.555	0	열 3	0.9884	0.5477	0

표 3.3: 모집단 II의 행(열) 간의 거리

	행 1	행 2	행 3	행 4		열 1	열 2	열 3	열 4
행 1	0	1.194	1.379	0.991	열 1	0	0.938	1.362	1.476
행 2	1.194	0	1.418	1.523	열 2	0.938	0	0.492	1.312
행 3	1.379	1.418	0	0.583	열 3	1.362	0.492	0	1.169
행 4	0.991	1.523	0.583	0	열 4	1.476	1.312	1.169	0

표 3.4: 모집단 I의 행(열) 간 거리의 포함확률

	명목확률 0.9			명목확률 0.95		
	$n_{..} = 200$	$n_{..} = 300$	$n_{..} = 500$	$n_{..} = 200$	$n_{..} = 300$	$n_{..} = 500$
행(열)d_{12}	.902(.901)	.903(.907)	.906(.906)	.952(.949)	.955(.958)	.951(.950)
행(열)d_{13}	.894(.899)	.903(.914)	.884(.905)	.949(.951)	.951(.958)	.943(.950)
행(열)d_{23}	.895(.910)	.884(.885)	.923(.906)	.952(.959)	.948(.945)	.962(.962)

표 3.5: 모집단 II의 행(열) 간 거리의 포함확률

	명목확률 0.9			명목확률 0.95		
	$n_{..}=200$	$n_{..}=300$	$n_{..}=500$	$n_{..}=200$	$n_{..}=300$	$n_{..}=500$
행(열)d_{12}	.885(.880)	.886(.882)	.881(.877)	.936(.942)	.940(.937)	.934(.933)
행(열)d_{13}	.926(.912)	.917(.920)	.910(.890)	.962(.969)	.955(.959)	.951(.947)
행(열)d_{14}	.880(.905)	.884(.905)	.866(.902)	.940(.954)	.945(.952)	.929(.952)
행(열)d_{23}	.894(.887)	.897(.886)	.882(.855)	.940(.940)	.944(.936)	.946(.917)
행(열)d_{24}	.954(.946)	.929(.946)	.908(.918)	.977(.980)	.966(.972)	.952(.958)
행(열)d_{34}	.878(.883)	.869(.888)	.859(.876)	.932(.939)	.922(.944)	.928(.928)

표 3.6: Greenacre 자료의 좌표점간의 거리에 대한 95% 신뢰구간

	붓스트랩 백분위 신뢰구간					
	표본거리	하한값	상한값	표본거리	하한값	상한값
행 d_{12}	0.329	0.080	0.997	행 d_{13} 0.364	0.093	1.066
행 d_{14}	0.391	0.096	0.940	행 d_{15} 0.304	0.083	1.028
행 d_{23}	0.681	0.159	1.188	행 d_{24} 0.302	0.082	0.903
행 d_{25}	0.562	0.117	1.199	행 d_{34} 0.617	0.299	0.988
행 d_{35}	0.201	0.074	0.729	행 d_{45} 0.435	0.109	0.888
열 d_{67}	0.522	0.183	0.904	열 d_{68} 0.591	0.269	0.943
열 d_{69}	0.707	0.284	1.138	열 d_{78} 0.165	0.047	0.527
열 d_{79}	0.391	0.081	0.954	열 d_{89} 0.227	0.063	0.796

4. 좌표축 설명력의 표본분포

대응분석의 결과를 저차원으로 표현할 경우 각 좌표축의 설명력은 해당하는 고유값들에 의해 표현되며 이에 대한 통계적추론 역시 매우 의미가 있다. 대응분석에 있어서 좌표축의 설명력은 조정된 대응행렬 $P-\mathbf{rc}^T$ 의 특이치분해를 이용하여 다음과 같이 표현한다. 즉, 조정된 대응행렬의 k번째 고유값을 μ_k라 할때 $(\mu_k)^2=\lambda_k$ 라 하자. 이때, 각 축의 설명력은

$$Y_1=\lambda_1/\sum_{k=1}^{I}\lambda_k : \text{첫번째 축의 설명력}$$

$$Y_2=\lambda_2/\sum_{k=1}^{I}\lambda_k : \text{두번째 축의 설명력}$$

$$Y=(\lambda_1+\lambda_2)/\sum_{k=1}^{I}\lambda_k : \text{첫번째와 두번째 축의 공동설명력}$$

이다. 그러므로 앞에서 언급된 $P^*-\mathbf{r}^*\mathbf{c}^{*T}$에 대한 특이치분해를 통해 구한 붓스트랩 고유값 $\mu_1^*\geq\cdots\geq\mu_I^*>0$ 을 사용하여 각 축의 설명력에 대한 추론을 시도하여 보자. 이 경우에는 고유값만이 관심의 대상이므로 주좌표를 구할 때와 달리 식별의 문제가 없다.

이제, $(\mu_k^*)^2=\lambda_k^*$ 이라 하면

$$Y_1^*=\lambda_1^*/\sum_{k=1}^{I}\lambda_k^* : \text{첫번째 축의 설명력}$$

$$Y_2^*=\lambda_2^*/\sum_{k=1}^{I}\lambda_k^* : \text{두번째 축의 설명력}$$

$$Y^*=(\lambda_1^*+\lambda_2^*)/\sum_{k=1}^{I}\lambda_k^* : \text{첫번째와 두번째 축의 공동설명력}$$

으로 표현할 수 있고 이제 각 축의 설명력에 대한 통계적 추론은 고유값의 붓스트랩 분포를 이용하여 가능할 수 있다. 가령, 첫번째 설명력(첫번째 고유값의 함수)의 표준오차 또는 신뢰구간은 Y_1^*의 붓스트랩 분포를 통해서 구할 수 있다. 이제 3절의 표3.1의 모집단 I을 이용하여 포함확률의 관점에서 이러한 붓스트랩 신뢰구간의 타당성을 확인하여보자. 가상모집단 I 에서의 1축, 2축의 설명력은 $Y_1=0.8033$, $Y_2=0.1976$ 그리고 1축, 2축의 고유값은 각각 $\lambda_1=0.1944$, $\lambda_2=0.0476$ 이 된다. 붓스트랩 신뢰구간 계산을 하기 위해서 붓스트랩 반복은 $B=500$이고 총 독립반복 시행횟수는 1000으로 하였다.

표4.1를 보면 좌표축 설명력의 신뢰도가 각각 90%와 95%인 붓스트랩 신뢰구간을 이용한 포함확률이 표본의 크기 n..가 증가함에 따라 명목 신뢰수준에 근사하고 있는 것을 볼 수 있다. 한편, 이제 앞에서 인용하였던 Greenacre(1984) 자료에서 제1축의 설명력은 0.8776(87.76%)이고 제 2축의 설명력은 0.1176(11.76%)으로 나타난다. 이에 대하여 제 1, 2축의 설명력에 대한 붓스트랩분포와 신뢰구간을 구한 결과가 그림4.1에 제시되어 있다. 제

표 4.1: 가상모집단 I의 축 설명력 Y_1, Y_2의 포함확률

	명목확률 0.9		명목확률 0.95	
	$n_{..} = 300$	$n_{..} = 500$	$n_{..} = 300$	$n_{..} = 500$
Y_1	0.873	0.885	0.923	0.941
Y_2	0.878	0.890	0.928	0.931

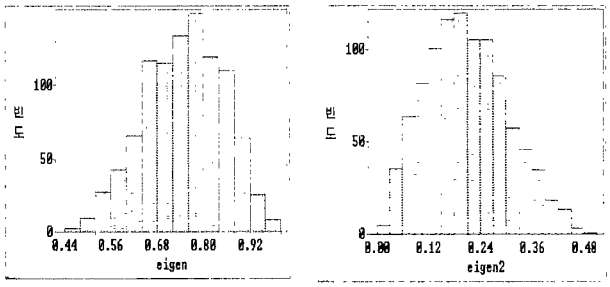

그림 4.1: 제1축 설명력(좌) 과 제2축 설명력(우)에 대한 붓스트랩 분포

1축의 90% 와 95% 백분위 신뢰구간은 각각 [.5686,.9088], [.5397,.9323] 이고 제 2축의 90% 와 95% 백분위 신뢰구간은 [.0671,.3812], [.0471,.4062] 로 나타났다.

5. 결론

대응분석에 사용되는 프로화일 또는 고유값등에 대한 전통적인 분포이론은 매우 미진하여 대응분석은 탐색적인 면만이 강조되어왔다. 본연구에서는 이차원 분할표 자료분석에 사용되는 대응분석에 붓스트랩 방법의 사용가능성을 모의실험을 통해 살펴보았다. 그 결과 좌표점의 변이와 좌표점간의 거리에 대한 통계적 추론을 붓스트랩방법으로 해결할 수 있다는 것을 보였고 또한, 좌표축의 설명력에 대한 붓스트랩 신뢰구간의 포함확률도 일치성을 가짐을 볼 수 있었다. 따라서, 대응분석에서도 자료에 근거한 붓스트랩 방법의 활용이 유익할 것으로 기대되며 그에 대한 이론연구도 의미가 있을 것으로 생각된다.

참고문헌

[1] Balbi, S. (1992). On stability in nonsymmetrical correspondence analysis using bootstrap, *Statistica Applicata*, **4**, 543-552.

[2] Efron, B. (1979). Bootstrap Methods : Another Look at the Jackknife, *Annals of Statistics*, **7**, 1-26.

[3] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM Monograph.

[4] Greenacre, M. (1984). *Theory and Application of Correspondence Analysis*, London, Academic Press

[5] Reiczigel, Jeno (1996). Bootstrap tests in correspondence analysis, *Applied Stochastic Models and Data Analysis*, **12**, 107-117.

[6] Woodroofe,M. and Jhun, M. (1988). Singh's Theorem in the Lattice Case, *Statistics and Probability Letters*, **7**, 201-205.

[2001년 5월 접수, 2001년 8월 채택]

The Application of Bootstrap Methods for Correspondence Analysis

Chang Wan Kang [1] Daehak Kim [2] M. Jhun [3]

ABSTRACT

Correspondence analysis is a multivariate technique that displays the rows and columns of a two-way contingency table as points in low-dimensional spaces. The graphical display is used for exploration of the relationships between rows and columns of the given contingency table. However, the analysis gives the same display as far as they have the same row and column profiles no matter what the sample size is. By applying the bootstrap methods for the correspondence analysis, stability of the display proportion of the variance explained by each axis were studied. Monte Carlo simulation shows the validity of proposed bootstrap confidence intervals for the distance between rows and columns.

Keywords: Contingency table; Correspondence analysis; Bootstrap method; Statistical inference; Coverage probability.

1) Assistant Professor, Dept. of Computer Science and Statistics, Dongeui University.
E-mail: cwkang@hyomin.dongeui.ac.kr
2) Professor, Dept. of Statistical Information, Catholic University of Daegu.
E-mail: dhkim@cuth.cataegu.ac.kr
3) Professor, Dept. of Statistics, Korea University.
E-mail: jhun@korea.ac.kr