The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.3.485
28(3), 485–494

Applications of Bootstrap Methods for Canonical Correspondence Analysis

Hyeon-Seok Ko[a] · Myoungshic Jhun[b] · Hyeong Chul Jeong[c,1]

[a]Rural Development Administration; [b]Department of Statistics, Korea University;
[c]Department of Applied Statistics, University of Suwon

(Received March 10, 2015; Revised April 20, 2015; Accepted May 7, 2015)

Abstract

Canonical correspondence analysis is an ordination method used to visualize the relationships among sites, species and environmental variables. However, projection results are fluctuations if the samples slightly change and consistent interpretation on ecological similarity among species tends to be difficult. We use the bootstrap methods for canonical correspondence analysis to solve this problem. The bootstrap method results show that the variations of coordinate points are inversely proportional to the number of observations and coverage rates with bootstrap confidence interval approximates to nominal probabilities.

Keywords: canonical correspondence analysis, bootstrap method, the variation of coordinate point, the distance between coordinate points, the explanation power of axis, statistical inference

1. 서론

주성분분석(principal component analysis), 대응분석(correspondence analysis), 중복분석(redundancy analysis), 그리고 정준대응분석(canonical correspondence analysis) 등은 분할표 상의 행과 열의 관계를 시각적으로 보여주는 커다란 장점이 있다. 그러나 이러한 방법들은 동일 모집단의 자료가 주어짐에도 불구하고, 자료나 표본의 크기가 약간 변화함에 따라 좌표 위치가 크게 변화하는 경향이 있어, 주어진 자료만으로 모집단 정보를 구조적으로 파악하기에는 다소 어려움을 지니고 있다고 할 수 있다. 모집단 정보를 보다 쉽게 이해하기 위해서는 좌표 위치들의 변이를 파악할 수 있으면 되는데, 이를 위해 우리는 붓스트랩 방법을 고려할 수 있다.

Efron (1979)에 의해 제안된 붓스트랩은 기저분포에 대한 모수적 가정없이도 표본으로부터 재표본을 취하는 방법으로 표본분포를 추정할 수 있다. 이 방법은 수리적인 해석이 가능한 통계방법 뿐만이 아니라, 컴퓨터의 계산능력을 사용하여 좋은 근사값을 구함으로써 훨씬 복잡한 통계문제를 효과적으로 처리하는데 도움을 준다 (Jhun 등, 1997). 붓스트랩 방법의 점근적 이론은 Efron (1982), Woodroofe와 Jhun (1988) 등을 참고할 수 있다.

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2012003020).
[1]Corresponding author: Department of Applied Statistics, University of Suwon, Hwaseong, Gyeonggi 445-743, Korea. E-mail: jhc@suwon.ac.kr

한편, 대응분석에서 붓스트랩 방법 활용가능성은 Greenacre (1984)에 의해 언급되었으며, Kang 등 (2001) 등은 대응분석에서 좌표점의 변이와 좌표점 간의 거리에 대한 통계적 추론을 붓스트랩 방법으로 해결할 수 있음을 보이고, 좌표축의 설명력에 대하여 붓스트랩 신뢰구간을 사용한 포함확률이 명목확률에 거의 근사하고 있음을 모의실험으로 제시한 바 있다. 그런데, 생태학에서는 환경인자를 고려하기에 대응분석의 대안으로 정준대응분석이 활용되고 있다. 정준대응분석은 Ter Braak (1986)이 소개한 이후로 생태학에서 가장 많이 사용되는 서열화 방법 중의 하나로, 장소(sites)별 종(species) 출현 자료에 환경변수(environmental variables)를 추가적으로 관측하였을 때 사용할 수 있는 분석법이다. 이 방법은 대응분석에 비해 모집단의 정보를 평면상에 유사하게 재현하는 능력은 약간 떨어지지만, 종 및 장소를 환경변수와 같은 평면에 행렬도(biplot)로 나타낼 수 있어서 해석이 용이하다는 장점을 지니고 있다 (Ko 등, 2015). 그런데, 정준대응분석 역시 대응분석이 지니고 있는 좌표점의 설명력 등 여러 통계적 추론 등의 문제를 지니고 있는데, 정준대응분석에 대해서는 붓스트랩을 활용하여 환경과 종과의 관계에 대한 통계적 추론 문제를 개선하는 연구는 현재 시도된 바 없다. 따라서 본 연구에서는 Kang 등 (2001)이 시도한 대응분석에서의 붓스트랩 방법을 활용해 정준대응분석에서 좌표점의 변이, 좌표점 간의 거리, 그리고 좌표축의 설명력에 대해 통계적으로 추론 가능성을 살펴보고 통계적 의미를 확인한다.

본 연구의 2장에서는 붓스트랩 방법에 의한 정준대응분석의 행렬도의 변화를 살펴보았다. 3장에서는 정준대응분석의 표본좌표점의 붓스트랩 분포를 다루었다. 그리고 4장에서는 좌표축의 설명력에 대한 붓스트랩 표본분포를 고려하여 붓스트랩 추론 가능성을 다루었다.

2. 정준대응분석

정준대응분석을 위해 사용되는 종발현 자료는 n개의 장소와 q개의 종으로 구성되어 있으며, $Y_{n\times q} = (y_{ik})$로 표기한다. 여기서 Y는 이차원 분할표 자료로 종 자료이다. 또한, n개의 장소에 대한 p개의 환경인자를 지닌 다변량 자료를 환경자료라 하고, $X_{n\times p} = (x_{ij})$로 표기하기로 한다. 이제, 종 자료 Y를 $Q = D_r^{-1/2}(P - rc')D_c^{-1/2}$로 변환한다. 여기서, $r = (r_1, \ldots, r_n)' = (y_{1+}, \ldots, y_{n+})'/y_{++}$, $c = (c_1, \ldots, c_q)' = (y_{+1}, \ldots, y_{+q})'/y_{++}$, $D_r = \mathrm{diag}(r_1, \ldots, r_n)$, $D_c = \mathrm{diag}(c_1, \ldots, c_q)$, $P = Y/y_{++}$, $y_{i+} = \sum_k y_{ik}$, $y_{+k} = \sum_i y_{ik}$, $y_{++} = \sum_i \sum_k y_{ik}$이다. 정준대응분석은 대응분석과 같이 행 프로파일(row profile) 정보와 카이제곱거리(chi-square distance)를 이용한다는 점에서 유클리디안 거리를 이용하는 주성분분석, 중복분석 등과 다르다. 카이제곱 거리의 저변에 깔린 의미는 전체적으로 출현빈도가 낮은 종이 어느 장소에서 관찰되었다면, 관찰되기 어려운데도 불구하고 관찰되었기 때문에 더 높은 가중치를 주자는 것이다. 그리고 환경 자료 X를 Z로 변환한다. 여기서 Z의 원소 z_{ik}는 $\sum_i w_i x_{ik} = 0$, $\sum_i w_i x_{ik}^2 = 1$을 만족하도록 x_{ik}에서 변환되었으며, $w_i = y_{i+}/y_{++}$는 가중치이다. 이것은 하나의 종이 관찰되었을 때, 환경 자료도 같이 측정되었기 때문에, 동일 기준으로 환경 자료를 표준화하기 위함이다. 이처럼 종 자료와 환경 자료의 사전 변환작업이 끝나면, 환경 자료가 종 출현에 영향을 주는 정도를 반영해 주기 위해 Q 대신 D_r을 가중값으로 하는 가중 다중회귀분석을 실시하여 $\hat{Q} = D_r^{1/2}Z(Z'D_rZ)^{-1}Z'D_r^{1/2}Q$을 분석에 사용한다. 이는 종출현이 많은 장소에 가중치를 주기 위함이다. 그리고 특이값분해 $\hat{Q} = UD_\mu V'$을 통해 종좌표 $D_c^{-1/2}VD_\mu$, 장소좌표 $D_r^{-1/2}U$, 환경변수벡터 $Cor(D_r^{1/2}Z, U)$를 구하여 종, 장소, 그리고 환경의 관계를 행렬도에 나타낸다. 즉, 정준대응분석은 가중다중회귀와 대응분석이 결합된 분석이라 할 수 있다.

3. 표본좌표점 거리 및 환경 변수 벡터 각도의 표본분포

정준대응분석 결과를 시각적으로 보기 위해 행렬도에 표현 할 경우, 각 좌표점 간(장소 간 혹은 종 간)

또는 환경변수 벡터 간에 대한 통계적 추론이 보다 의미를 지니게 된다. 이것은 생태학적인 측면에서, 좌표점들끼리 더 가까운 것들은 서로 간 생태학적인 유사성을 보이기도 때문이다.

이차원상에 근사되어진 임의의 두 장소 또는 종의 좌표점 $i(a_i, b_i)$와 $k(a_k, b_k)$의 거리는 식 (3.1)로 계산할 수 있다.

$$d_{ik} = \sqrt{(a_i - a_k)^2 + (b_i - b_k)^2}. \tag{3.1}$$

정준대응분석의 또 다른 요소인 환경변수는 장소와 종의 좌표점과는 달리 방향성을 가지고 있는 벡터이기 때문에 거리가 아닌 각도로 표시할 수 있다. 환경변수 벡터 $\vec{e}_i(a_i, b_i)$와 $\vec{e}_k(a_k, b_k)$ 간 각도는 식 (3.2)로 계산된다.

$$\theta_{ik} = \arccos\left(a_i/\sqrt{a_i^2 + b_i^2}\right) - \arccos\left(a_k/\sqrt{a_k^2 + b_k^2}\right). \tag{3.2}$$

그런데, d_{ik}와 θ_{ik}는 두 장소(종) 사이의 실제 거리와 두 환경변수 벡터 사이의 실제 각도가 아니기 때문에, 그에 대한 변이를 모르는 상태에서는 앞의 두 값을 확증적 자료분석에 그대로 사용하기는 다소 문제를 지닌다. 이제 d_{ik}와 θ_{ik}에 대한 변이를 추정하기 위해 붓스트랩 방법을 사용하기로 한다. 표본좌표점의 거리를 추정하는 붓스트랩 절차는 다음과 같다. 한편, 표본좌표점의 방향성 식별에 대한 붓스트랩 방법은 Balbi (1992)와 Kang 등 (2001)을 참고할 수 있다.

단계 1: 모집단 종 자료의 모집단 비율 P를 정의하고 환경 자료 X를 다변량정규분포에서 생성한다. 본 실험에서 사용한 다항분포 모비율(P)과 다변량정규분포의 평균과 분산은 다음과 같다.

$$P = \begin{pmatrix} 0.20 & 0.05 & 0.05 \\ 0.10 & 0.20 & 0.10 \\ 0.05 & 0.10 & 0.15 \end{pmatrix}, \qquad \mu = (10, 30, 50), \qquad \Sigma = \begin{pmatrix} 10 & 5 & 10 \\ 5 & 20 & 15 \\ 10 & 15 & 30 \end{pmatrix}.$$

단계 2: 해당 자료에 대한 행좌표(F), 열좌표(G), 환경변수 벡터(E), 두 행(열) 좌표간의 거리 d_{ik} 및 환경변수 벡터간의 각도를 계산한다.

단계 3: 모집단 P에서 추출된 표본 Y에 기초하여 붓스트랩 표본 Y^*를 생성한다.

단계 4: 붓스트랩 표본에 기초하여 행 좌표 F^*, 열 좌표 G^*, 환경변수 벡터 E^*, 두 행(열) 좌표간의 거리 d_{ik}^* 및 환경변수 벡터간의 각도 θ_{ik}^*를 계산한다.

단계 5: 붓스트랩 반복($B = 500$)으로부터, 각 통계량의 $100(1 - \alpha/2)\%$ 신뢰구간을 계산한 후, 모표본으로 계산된 거리의 포함여부를 판정한다.

단계 6: (단계 3)에서 (단계 5)를 $M = 1000$회 반복하여 포함확률을 계산한다.

Table 3.1은 표본 장소와 종 간 거리와 환경변수 벡터 간 각도의 신뢰도가 90%, 95%인 붓스트랩 신뢰구간에 대한 포함확률을 보여준다. 여기서 포함확률이 명목확률에 거의 근사하고 있는 것을 볼 수 있으며, 이로부터 붓스트랩 거리분포의 타당성을 확인할 수 있다. 한편, 장소와 종 간 거리에 영향을 받기에 분산의 크기는 포함확률의 변화에 큰 영향을 끼치지는 못한다.

4. 좌표축 설명력의 표본분포

정준대응분석의 결과를 행렬도에 표현할 경우, 각 좌표축의 설명력은 해당하는 고유값들의 크기로 측정된다. 정준대응분석에 있어서 좌표축의 설명력은 $\hat{Q} = UD_{\sqrt{\lambda}}V'$의 특이값분해를 사용하여 제 1축 설명

Table 3.1. Coverage rates of population distance between rows (columns) and population angles between environmental variables vectors

	$n_{..} = 200$	$n_{..} = 300$	$n_{..} = 500$
	nominal probability 0.90 (row, column, environment)		
d_{12}	(0.897, 0.883, 0.904)	(0.903, 0.897, 0.895)	(0.890, 0.896, 0.895)
d_{13}	(0.905, 0.891, 0.893)	(0.895, 0.898, 0.903)	(0.898, 0.897, 0.896)
d_{23}	(0.902, 0.884, 0.904)	(0.907, 0.885, 0.905)	(0.887, 0.892, 0.890)
	nominal probability 0.95 (row, column, environment)		
d_{12}	(0.943, 0.939, 0.949)	(0.953, 0.945, 0.955)	(0.947, 0.944, 0.947)
d_{13}	(0.953, 0.939, 0.937)	(0.958, 0.945, 0.949)	(0.947, 0.949, 0.947)
d_{23}	(0.947, 0.941, 0.950)	(0.952, 0.948, 0.947)	(0.943, 0.952, 0.949)

Table 4.1. Coverage rates of explanation power of the population axis

	$n_{..} = 200$	$n_{..} = 300$	$n_{..} = 500$
		nominal probability 0.90	
first axis	0.911	0.907	0.911
second axis	0.912	0.904	0.909
		nominal probability 0.95	
first axis	0.954	0.958	0.943
second axis	0.955	0.955	0.942

력은 $\lambda_1/\sum\lambda_i$, 제 2축 설명력은 $\lambda_2/\sum\lambda_i$이다. 그러나 고유값들의 변이를 모르기 때문에 좌표축 설명력에 대한 추론에 어려움이 있다. 그러므로 각 축의 설명력에 대한 통계적 추론을 위해 고유값의 붓스트랩 분포를 활용하기로 한다. 좌표축 설명력을 유도하는 붓스트랩 절차는 다음과 같다.

단계 1: 모집단 종 자료의 모집단 비율 P를 정의하고 환경 자료 X를 다변량정규분포에서 생성한다. 본 실험에서 사용한 다항분포 모비율(P)과 다변량정규분포의 평균과 분산은 다음과 같다.

$$P = \begin{pmatrix} 0.20 & 0.05 & 0.05 \\ 0.10 & 0.20 & 0.10 \\ 0.05 & 0.10 & 0.15 \end{pmatrix}, \qquad \mu = (10, 30, 50), \qquad \Sigma = \begin{pmatrix} 10 & 5 & 10 \\ 5 & 20 & 15 \\ 10 & 15 & 30 \end{pmatrix}.$$

단계 2: 해당 자료에 대한 전처리 과정을 거쳐 $\hat{Q} = UD_{\sqrt{\lambda}}V'$의 특이값 분해를 실시한다. 그리고, 1축과 2축의 설명력을 계산한다.

제 1축 설명력: $\lambda_1/\sum\lambda_i$.

제 2축 설명력: $\lambda_2/\sum\lambda_i$.

단계 3: 모집단 P에서 추출된 표본 Y에 기초하여 붓스트랩 표본 Y^*를 생성한다.

단계 4: 붓스트랩 표본으로 유도된 $\hat{Q}^*$를 특이값 분해 한다. 즉, $\hat{Q}^* = U^*D_{\sqrt{\lambda^*}}V^{*\prime}$이다. 이를 기초로, 제 1축과 2축에 대한 붓스트랩 설명력을 계산한다.

제 1축 설명력: $\lambda_1^*/\sum\lambda_i^*$.

제 2축 설명력: $\lambda_2^*/\sum\lambda_i^*$.

단계 5: 붓스트랩 반복($B = 500$)으로부터, 각 통계량의 $100(1-\alpha/2)\%$ 신뢰구간을 계산한 후, 모표본으로 계산된 각 축 설명력의 포함여부를 판정한다.

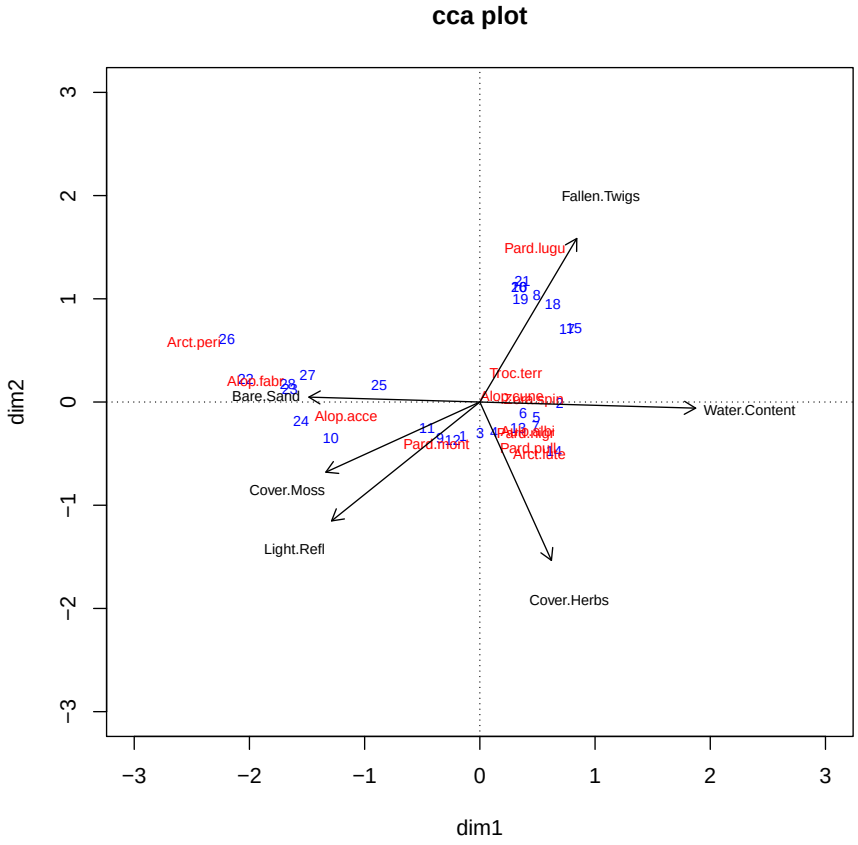

Figure 5.1. Biplot of canonical correspondence analysis for the spider data.

단계 6: (단계 3)에서 (단계 5)를 $M = 1000$회 반복하여 포함확률을 계산한다.

Table 4.1는 각 축에 대한 붓스트랩 방법으로 계산한 신뢰수준을 보여준다. 좌표축 설명력의 신뢰도가 가각 90%, 95%인 붓스트랩 신뢰구간을 사용한 포함확률이 표본의 크기가 증가함에 따라 명목 신뢰수준에 근사하고 있음을 볼 수 있다. 이러한 결과는 정준대응분석의 좌표축 설명력에 대해 붓스트랩 추론이 타당성을 지니고 있음을 보여준다고 할 수 있다.

5. 사례분석

5.1. 분석 자료

Ter Braak (1986)의 거미 자료는 1975년에 네덜란드의 모래언덕 28개 장소에서 12종의 거미 출현 빈도와 6개의 환경변인을 조사한 것이다. 12종의 거미 출현 자료는 높은 빈도에 의한 분산을 줄이기 위해 제곱근 변환 후 정수 부분만 선택하였고, 값이 9보다 큰 경우는 9로 표기하였다. 6개의 환경변인은 수분함량(water content), 사막화 비율(bare sand), 이끼 분포 비율(cover moss), 빛 반사 비율(light reflection), 떨어진 나뭇가지 비율(fallen twig), 엽채류 분포 비율(cover herbs)이며, 로그변환 후 10개

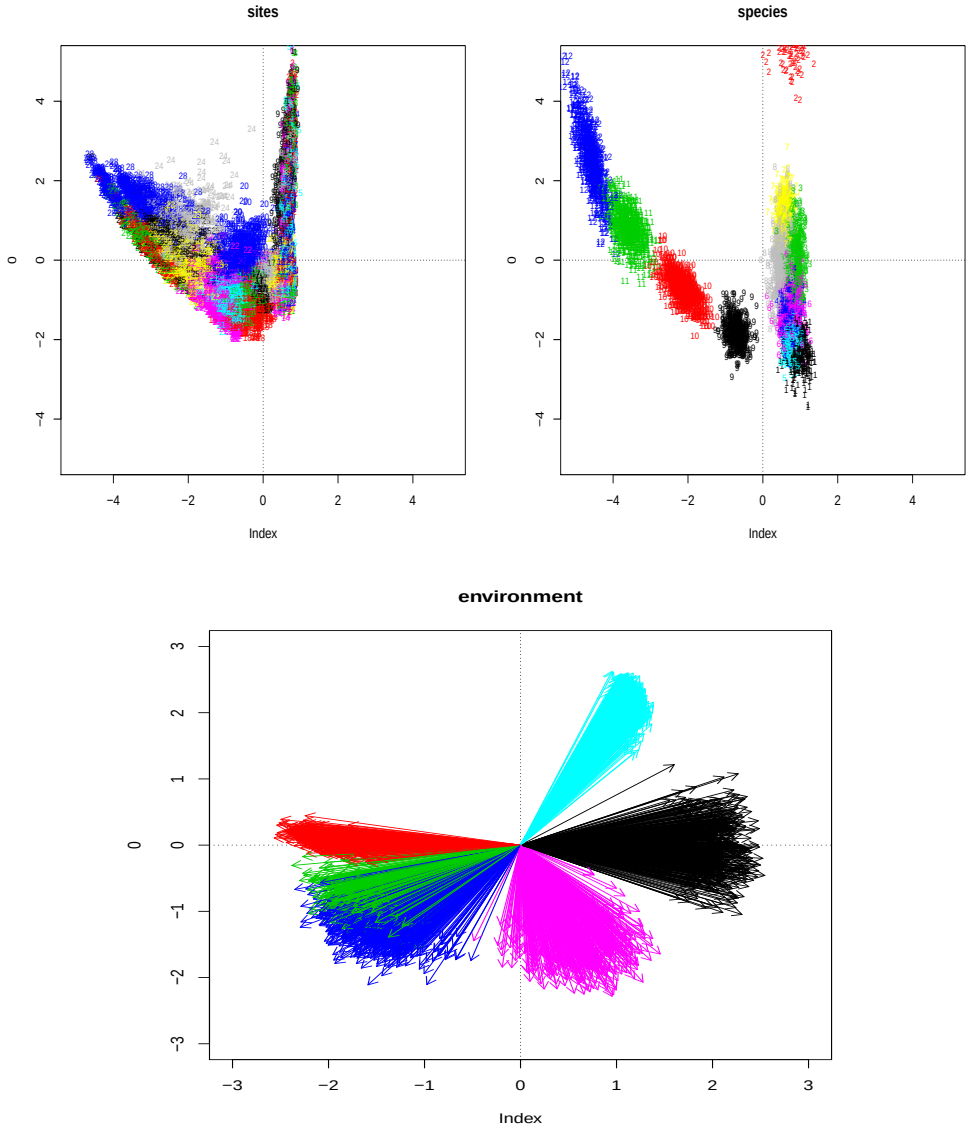

Figure 5.2. Bootstrap biplot of canonical correspondence analysis for the spider data.

의 동일 크기로 나누고 0에서 9까지의 값을 주었다. 본 절에서는 Ter Braak (1986)의 자료를 사용하여 붓스트랩 행렬도 활용을 소개하기로 한다.

5.2. 붓스트랩 행렬도에 근거한 외적 안정성

정준대응분석은 환경변수의 변화에 따른 종과 장소의 관계를 시각적인 행렬도로 표현할 수 있다. 다변량자료에 대한 행렬도에 대해서는 Gabriel (1971), Lebart 등 (1984)을 참고할 수 있다. 정준대응분석

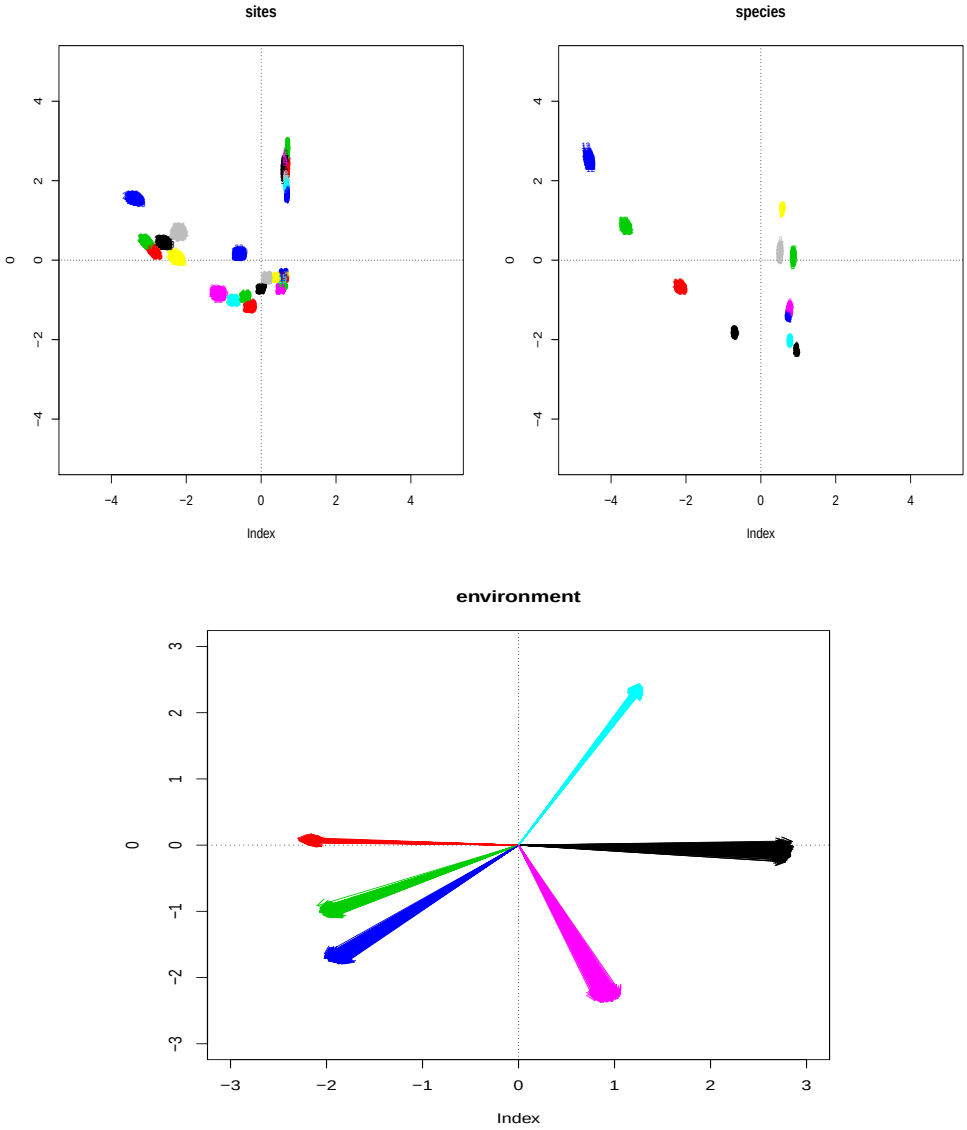

Figure 5.3. Bootstrap biplot of canonical correspondence analysis for the spider data (100 magnification).

의 행렬도는 종 발현자료에 대한 종과 장소의 이차원 행렬도 위에 환경인자 벡터가 동일 공간에 표현되는 구조로 나타난다. Figure 5.1은 Ter Braak (1986)의 거미 자료에 대한 정준대응분석 행렬도인데, 종, 장소 그리고 환경변수가 같이 표현되어 있음을 볼 수 있다. 한편, Figure 5.1의 자료에 대한 행렬도 결과는 뒤의 붓스트랩 분석의 원형 역할을 담당하게 된다. 그런데, 이 분석은 전체 자료 중 일부 표본에 대한 결과로서 표본이 바뀔 때마다 행(열) 프로파일이 변화하여 서로 다른 결과를 보이는 단점이 있다. 이러한 변화는 생태학적인 변인들 간의 관계에 대한 일관된 해석을 어렵게 하기 때문에, 각 요인을 이차원 공간의 하나의 좌표로 나타내기보다는 분포로 표현하는 것이 자료의 기저를 이해하는데 더 유용하리

라 생각된다.

포지션닝 좌표를 분포로 표현하기 위해서는 붓스트랩 방법을 활용할 수 있다. 또한, 붓스트랩을 사용하여 표본의 크기와 분포의 변이를 연관시킬 수 있다. 다음은 정준대응분석의 외적안정성을 높이기 위한 붓스트랩 절차이다.

단계 1: Y의 빈도에 기초하여 다항분포로부터 붓스트랩 표본 Y^*를 생성한다.

단계 2: 자료 Y^*와 X를 각각 다음의 $Q^* = D_{r^*}^{-1/2}(P^* - r^*c^{*\prime})D_{c^*}^{-1/2}$와 $Z^* = w_i^* X$로 변환한다.

단계 3: 가중회귀분석으로 $\hat{Q}^* = D_{r^*}^{1/2} Z^* (Z^{*\prime} D_{r^*} Z^*)^{-1} Z^{*\prime} D_{r^*}^{1/2} Q^*$를 계산한다.

단계 4: $\hat{Q}^* = U^* D_{\sqrt{\lambda}^*} V^{*\prime}$으로 특이값 분해 한다.

단계 5: 행과 열 좌표 및 환경변수 벡터를 계산한다.

$$\text{행: } F^* = D_{r^*}^{-1} P^* G D_{\sqrt{\lambda}}^{-2},$$
$$\text{열: } G^* = D_{c^*}^{-1} P^{*\prime} F D_{\sqrt{\lambda}}^{-2},$$
$$\text{환경변수: } E^* = \operatorname{cor}\left(D_{r^*}^{\frac{1}{2}} Z^*, U^* \right).$$

단계 6: 단계 1부터 단계 5까지를 붓스트랩 반복 ($B = 500$)한다.

Figure 5.2는 거미 자료에 대해 붓스트랩 알고리즘을 사용하여 장소, 종 그리고 환경변수 벡터에 대한 행렬도를 그린 것으로, 행렬도를 통해 각 좌표와 벡터의 산포를 파악 할 수 있다. 그런데 Figure 5.2에 나타난 분포의 변이가 상대적으로 큼을 볼 수 있다.

Figure 5.3은 거미 자료에 100배를 한 후 붓스트랩 방법으로 정준대응분석의 장소와 종 좌표, 그리고 환경변수 벡터의 행렬도를 그린 것이다. 이 결과는 각 포지셔닝 된 좌표와 벡터의 응집력을 확실히 보여주는데, 이를 통해 종과 환경관의 관계를 보다 명확히 볼 수 있으며, 서로 간의 거리 추정을 통해 변인들 간의 생태학적인 유사성을 파악하는데 도움을 준다. 또한, Figure 5.1에 근사되는 응집력을 볼 수 있다.

6. 결론

본 연구에서는 표본이 바뀔 때마다 서열화 방법의 결과가 달라지는 문제로 인해 발생하는 통계적 추론의 어려움을 해결하기 위해, 생태학에서 서열화 방법으로 널리 활용되는 정준대응분석에 붓스트랩 방법의 적용을 다루었다. 대응분석에 붓스트랩 방법을 적용하는 연구는 Kang 등 (2001)에서 참고할 수 있는데, 본 연구에서는 환경인자까지 고려하는 보다 복잡한 행렬도 구조하의 붓스트랩 가능성을 살펴본 연구로 의의를 둔다. 즉, 환경변수 벡터를 더 고려해야 하는 정준대응분석에서 붓스트랩 방법을 활용하여 자료의 외적안정성 및 표본좌표점 거리, 좌표축 설명력에 대한 표본분포를 살펴보았다. 분석결과 전체 관찰 자료수에 반비례하여 좌표점의 변이가 나타났으며, 붓스트랩 신뢰구간을 사용한 포함함률이 명목확률에 거의 근사함을 볼 수 있었다. 이를 통해 정준대응분석에서도 붓스트랩 방법으로 다양한 통계적 추론을 적용할 수 있음을 확인하였다.

References

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

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife, *Annals of Statistics*, **7**, 1–26.

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

Gabriel, K. R. (1971). The biplot display of matrices with the application to principal component analysis, *Biometrika*, **58**, 453–467.

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

Jhun, M., Jeong, H. C. and Jin, S. H. (1997). *The Understanding of Bootstrap Method*, Freeacademy.

Kang, C. W., Kim, D. and Jhun, M. (2001). The application of bootstrap methods for correspondence analysis, *The Korean Journal of Applied Statistics*, **14**, 401–413.

Ko, H. S., Jhun, M. and Jeong, H. C. (2015). A comparison study for ordination methods in ecology, *The Korean Journal of Applied Statistics*, **28**, 49–60.

Lebart, L., Morineau, A. and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*, Wiley, New York.

Ter Braak, C. J. F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis, *Ecology*, **67**, 1167–1179.

Woodroofe, M. and Jhun, M. (1988). Singh's theorem in the lattice case, *Statistics and Probability Letters*, **7**, 201–205.

정준대응분석에서 붓스트랩 방법 활용

고현석[a] · 전명식[b] · 정형철[c,1]

[a]농촌진흥청, [b]고려대학교 통계학과, [c]수원대학교 통계정보학과

(2015년 3월 10일 접수, 2015년 4월 20일 수정, 2015년 5월 7일 채택)

요 약

정준대응분석은 생태학에서 장소, 종 그리고 환경변수의 관계를 시각적으로 보기 위해 가장 많이 사용되는 서열화 방법 중의 하나이다. 그런데 이 방법은 표본이 바뀔 때마다 분석결과가 달라지기 때문에 종 간의 생태학적 유사성에 대한 일관된 해석을 어렵게 한다. 본 연구에서는 이러한 문제점을 해결하기 위해 정준대응분석에 붓스트랩 방법을 활용하였다. 이를 통해 전체 관찰 자료수에 반비례하여 좌표점의 변이가 나타나고, 붓스트랩 신뢰구간을 사용한 포함확률이 명목확률에 근사함을 확인하였다.

주요용어: 정준대응분석, 붓스트랩 방법, 좌표점의 변이, 좌표점의 거리, 좌표축의 설명력, 통계적 추론

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2012-003020).

[1]교신저자: (445-743) 경기도 화성시 봉담읍 와우리 산 2-2, 수원대학교 통계정보학과. E-mail: jhc@suwon.ac.kr