The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.1.049
28(1), 49–60

A Comparison Study for Ordination Methods in Ecology

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 October 22, 2014; Revised December 23, 2014; Accepted January 20, 2015)

Abstract

Various kinds of ordination methods such as correspondence analysis and canonical correspondence analysis are used in community ecology to visualize relationships among species, sites, and environmental variables. Ter Braak (1986), Jackson and Somers (1991), Parmer (1993), compared the ordination methods using eigenvalue and distance graph. However, these methods did not show the relationship between population and biplot because they are only based on surveyed data. In this paper, a method that measures the extent to show population information to biplot was introduced to compare ordination methods objectively.

Keywords: Distance graph, principal component analysis, correspondence analysis, redundancy analysis, canonical correspondence analysis, canonical ordination analysis, singular value decomposition.

1. 서론

군생태학(community ecology)에서는 종(species), 장소(sites) 그리고 환경변수(environmental variables) 사이의 관계를 규명함으로써, 어느 한 종이 어떤 환경과 어떤 장소에 적합한지 살펴보고, 환경인자의 동적 변화에 따라 종의 분포가 어떻게 변화하는지 예측하는데 높은 연구관심을 두고 있다. 이러한 목적을 위해 군생태학에서는 서열화방법(ordination method)을 사용하는데, 서열화방법은 환경변수의 유무에 따라 서열분석(ordination analysis) 및 정준서열분석(canonical ordination analysis)으로 구분된다.

서열분석은 종별 분포를 대표할 수 있는 일정한 크기의 표본 추출된 장소에서 종별 출현빈도를 조사한 자료를 이용하여 종과 장소의 연관성을 보고자 하는 분석으로, Bray와 Curtis (1957)가 극점서열분석(polar ordination)을 소개한 후 군생태학에서 널리 활용되고 있다. Ter Braak (1986)의 정준대응분석(canonical correspondence analysis)이 제시되기 이전의 서열분석에서는 환경 정보가 없는 종-발현 자료만으로 가상의 잠재변수를 도출한 후 이를 간접적 환경인자로 놓고 종 출현과 환경과의 관계를 규명하고자 하였다. 이를 간접 서열화(indirect ordination) 분석이라 한다. 그런데, 이런 잠재적 환경 인

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

자는 여러 환경변수들이 혼재 된 결과로 보이기에 자료 해석에 어려움이 존재한다. 간접 서열화 분석으로는 주성분분석(principal component analysis; PCA), 대응분석(correspondence analysis; CA) 등을 들 수 있다.

정준서열분석은 서열분석에 사용된 자료에 환경인자에 대한 추가적인 정보가 주어질 때 사용할 수 있는 방법이다. 이 분석법은 서열분석에 비해 조사된 자료의 종과 장소의 관계를 그림 상에 완전 유사하게 나타낼 수는 없지만, 환경변수와의 연관관계를 직접적으로 표현할 수 있기 때문에 매우 유용한 분석으로 알려진다. 정준서열분석에는 중복분석(redundancy analysis; RDA), 정준대응분석(canonical correspondence analysis; CCA) 등이 있다. 특히, Ter Braak (1986)의 정준대응분석은 환경변화에 따른 종분포를 잘 나타내주기 때문에 최근까지 가장 널리 사용되고 있는 서열화방법이다. Ter Braak (1986), Jackson과 Somers (1991), Parmer (1993) 등은 서열분석의 특징을 비교하기 위해 다양한 척도 및 방법들을 사용하였는데, 그중에서 많이 언급된 비교방법은 조사된 자료를 행렬도(biplot)에 어느 정도 근사하게 재현하는 정도를 표현하는 고유값(eigenvalue)들의 비교였다. 그런데, 이러한 고유값 비교는 '모집단 전체에 대해서가 아닌 조사된 자료에 대해서만 어느 정도 행렬도에 재현되는가?' 하는 정도만을 평가는 것이다. 즉, 모집단의 일부 자료만을 조사하기 때문에, 행렬도에 모집단 정보를 보다 정확하게 표현하는지를 비교하기 위해서는 새로운 평가방법이 필요하다. 본 연구의 2장에서는 대표적인 서열화방법들을 소개하고 3장에서 이러한 서열화 방법들을 비교하기 위한 기존의 방법들을 언급하고자 한다. 그리고 4장에서는 서열화방법에 의해 모집단 정보의 행렬도 근사정도를 측정할 수 있는 새로운 지표를 소개하고 모의실험을 통해 서열화방법들을 비교하였다. 끝으로 5장에서 결론을 내리고자 한다.

2. 대표적인 서열화방법 및 분석 자료

생태학에서 서열화방법이란 좌표 상에 종 또는 장소를 어떤 순서에 따라 위치를 정하는 것으로, 분석에 사용되는 자료에 환경 정보가 있는가 없는가에 따라 서열분석(비정준서열분석)과 정준서열분석으로 구분된다. 그런데, 고차원의 다변량 자료 서열 구조를 삼차원 이하의 저차원에 표현하여 낮은 차원에서 서열구조를 파악하기 위해서는 차원축소 기법이 요구된다. 이를 위해 서열분석에서 특이값분해(singular value decomposition; SVD)를 사용할 수 있다. 본 분석에서 사용되는 종-발현 자료는 n개의 장소와 q개의 종으로 구성되어 있으며 $Y_{n\times q} = (y_{ik})$로 표기한다. 각 칸은 i번째 장소에서 k번째 종의 출현빈도를 의미한다. 한편, 정준서열분석에서는 n개의 장소와 p개의 환경변수로 구성된 환경자료가 추가적으로 주어지는데, 이는 $X_{n\times p} = (x_{ij})$로 표기한다. 생태학의 종-발현 자료에 대해서는 Jeong (2012)을 참고할 수 있다. 한편, 자료를 변환하고 가중치를 부여하는 방법에 따라 여러 유형의 서열분석이 존재하는데 이를 2.1절에서 언급하기로 한다.

2.1. 대표적인 서열화방법

2.1.1. 주성분분석 서열분석에서 초기에 널리 사용된 분석법은 주성분분석이다. 주성분분석에서는 자료를 가장 잘 축약할 수 있도록 기존 변수에 대해 선형결합을 실시하여, 서로 직교 기저인 주성분변수를 만든다. 그리고, 정보손실을 최소화하면서 설명이 간결할 수 있도록 소수의 주성분변수를 이용하여 자료를 설명하고 시각화를 시도한다. 우선, 종 발현 자료 Y를 발현 빈도나 단위의 영향력을 제거하기 위해 각 종 변수의 평균과 표준편차를 이용하여 식 (2.1)과 같이 표준화 한다.

$$Y_s = \frac{(y_{ik} - \bar{y}_k)}{s_{yk}}, \tag{2.1}$$

여기서 $\bar{y}_k = \sum_{i=1}^{n} y_{ik}/n$, $s_{yk} = \sqrt{\sum_{i=1}^{n}(y_{ik} - \bar{y}_k)^2/(n-1)}$ 이다. 그리고 특이값분해 $Y_s = UD_\mu V'$를 통해 종좌표 V 및 장소좌표 UD_μ를 구한다. 주성분분석의 서열화 과정은 Legendre와 Legendre (2012)을 참고할 수 있다.

2.1.2. 대응분석 Hill (1973)의 상호평균법(reciprocal averaging)이 생태학에 소개된 후 대응분석은 서열분석의 도구로서 가장 널리 사용되고 있다. Hill의 상호평균법은 Hayashi의 수량화 3법, Benzecri (1973)의 대응분석과 같은 방법이다 (Kim과 Jeong, 2013). 대응분석은 행프로파일(row profile) 정보와 카이제곱거리(chi-square distance)를 이용한다는 점에서 유클리디안 거리(euclidean distance)를 이용하는 주성분분석과는 차이가 있다. 생태통계학에서 카이제곱거리는, 대부분의 장소에서 출현빈도가 낮은 특정한 종이 어느 특정 장소에서 집중적으로 관찰되었다면, 해당 종의 관찰(혹은 발현)이 매우 드문 현상임에도 불구하고 특정 장소에서 많이 관찰되었기에 그 장소에 더 높은 가중치를 준다는 의미를 지니고 있다. 이 방법은 식 (2.2)와 같이 종-발현 자료 Y를 Q로 변환한다.

$$Q = D_r^{\frac{1}{2}} D_r^{-1} \left(P - rc'\right) D_c^{-\frac{1}{2}} = D_r^{-\frac{1}{2}} \left(P - rc'\right) D_c^{-\frac{1}{2}}, \tag{2.2}$$

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. 그리고 특이값분해 $Q = UD_\mu V'$를 통해 종좌표 $D_c^{-1/2} V D_\mu$ 및 장소좌표 $D_r^{-1/2} U$를 유도한다.

2.1.3. 중복분석 중복분석은 주성분분석을 활용하는 정준서열분석의 일종이다. 종-발현 자료에 추가적으로 환경정보가 존재할 때 사용할 수 있다. 먼저 주성분분석처럼 종-발현자료를 Y를 Y_s로, 환경자료 X를 X_s로 표준화한다. 그리고 다중회귀를 통해 Y_s의 적합값 $\hat{Y} = X_s(X_s'X_s)^{-1}X_s'Y_s$를 구한다. 그리고 특이값분해 $\hat{Y} = UD_\mu V'$를 통해 종좌표 V, 장소좌표 UD_μ, 환경변수벡터 $\mathrm{Corr}(X_s, U)$를 계산한다.

2.1.4. 정준대응분석 Ter Braak (1986)의 정준대응분석은 생태학에서 가장 널리 사용되는 서열화방법 중 하나라 할 수 있다. 정준대응분석은 종-발현 정보에 더하여 해당 장소의 환경정보가 주어질 때 사용할 수 있는 대응분석의 일종으로, 서열화 성능은 대응분석에 비해 다소 떨어지지만 종 및 장소를 환경변수와 같이 나타낼 수 있어서 자료해석이 용이하다는 장점이 지니고 있다. 여기서 서열화 성능이란 조사된 자료의 종과 장소의 관계를 행렬도에 유사하게 나타내는 정도를 의미한다. 이 방법은 종-발현 자료 Y를 대응분석처럼 Q로 변환하고, 환경자료 X를 Z로 변환한다. 여기서 Z의 원소 z_{ik}는 $\sum_i w_i x_{ik} = 0$, $\sum_i w_i x_{ik}^2 = 1$를 만족하도록 x_{ik}로 부터 변환된다. 여기서, $w_i = y_{i+}/y_{++}$이다. 그리고 가중다중회귀를 통해 Q의 적합값 $\hat{Q} = D_r^{1/2} Z (Z' D_r Z)^{-1} Z' D_r^{1/2} Q$를 구한 후 특이값분해 $\hat{Q} = UD_\mu V'$를 통해 종좌표 $D_c^{-1/2} V D_\mu$, 장소좌표 $D_r^{-1/2} U$, 환경변수벡터 $\mathrm{Corr}(D_r^{1/2} Z, U)$를 계산한다.

2.2. 분석 자료

본 연구에서는 Ter Braak (1986)의 거미 자료를 사용하여 여러 방법들을 비교하기로 한다. Ter Braak (1986)의 거미 자료는 1975년에 네덜란드의 28개 모래언덕에서 12종 거미의 출현 정도를 관찰하고, 동시에 해당 28개 모래언덕의 여섯가지 환경인자를 측정한 자료이다. 12종의 거미 출현 자료에서 높은 빈

Table 3.1. Eigenvalues of four analysis

Dimension	Principal component analysis			Redundancy analysis			
	Eigenvalue	Proportion	Cumulative proportion	Eigenvalue	Proportion	Cumulative proportion	Adjusted cumulative
1	165.311	51.0	51.0	136.888	61.1	61.1	42.2
2	63.414	19.6	70.6	54.774	24.5	85.6	59.2
3	35.897	11.1	81.7	17.797	7.9	93.5	64.6
4	20.014	6.2	87.9	9.218	4.1	97.6	67.5
5	10.116	3.1	91.0	4.278	1.9	99.5	68.8
Dimension	Correspondence analysis			Canonical correspondence analysis			
	Eigenvalue	Proportion	Cumulative proportion	Eigenvalue	Proportion	Cumulative proportion	Adjusted cumulative
1	0.589	48.5	48.5	0.540	61.9	61.9	44.5
2	0.254	20.9	69.4	0.225	25.7	87.6	62.9
3	0.180	14.8	84.2	0.073	8.4	96.0	69.0
4	0.065	5.3	89.5	0.020	2.3	98.3	70.6
5	0.040	3.3	92.8	0.011	1.3	99.6	71.5

도를 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), 빛반사 비율(light reflection), 떨어진 나뭇가지 비율(fallen twig), 엽채류 분포 비율(cover herbs)이며 로그변환 후 10개의 동일 크기로 나누고 0에서 9까지의 값을 부여하였다.

3. 기존의 서열화방법 비교법

위에서 소개한 서열화방법들을 통해 종 및 장소의 좌표, 환경변수 벡터를 q차원까지 계산할 수 있으나, 시각화 측면에서는 삼차원까지만을 그림으로 표현할 수 있다. 그런데, 삼차원 그림은 보는 방향에 따라 서열화 결과가 다르게 해석될 여지가 존재함으로, 서열화 결과를 이차원 공간에 표현하는 것이 합리적이라 생각된다. 따라서 정보의 손실을 감수하고 2개의 축을 활용하여 그림으로 표현하는데, 표현된 이차원 그림의 성능을 평가하기 위한 평가 지표가 필요하다. 이에 대해서는 고유값과 이차원 그림상의 각 개체들 간의 거리가 주로 활용된다.

3.1. 고유값을 이용한 설명력 비교법

자료행렬 Y가 $UD_\mu V'$로 특이값분해 된다고 하면, $Y = \sum_{k=1}^{q} \sqrt{\lambda_k} u_k v_k'$로 쓸 수 있으며 s차원까지만 사용한다면 $Y_s = \sum_{k=1}^{s} \sqrt{\lambda_k} u_k v_k$가 된다. 따라서 자료행렬 s차원의 근사도를 $A_s = 1 - ||Y - Y_s||^2/||Y||^2$로 정의하면, A_s를 다음 식 (3.1)로 표현할 수 있다.

$$A_s = 1 - \frac{\left|\left| \sum_{k=s+1}^{q} \sqrt{\lambda_k} u_k v_k' \right|\right|^2}{\left|\left| \sum_{k=1}^{q} \sqrt{\lambda_k} u_k v_k' \right|\right|^2} = 1 - \frac{\sum_{k=s+1}^{q} \lambda_k}{\sum_{k=1}^{q} \lambda_k} = \frac{\sum_{k=1}^{s} \lambda_k}{\sum_{k=1}^{q} \lambda_k}. \tag{3.1}$$

이제, 고유값을 이용한 서열화방법들의 설명력을 살펴보자.

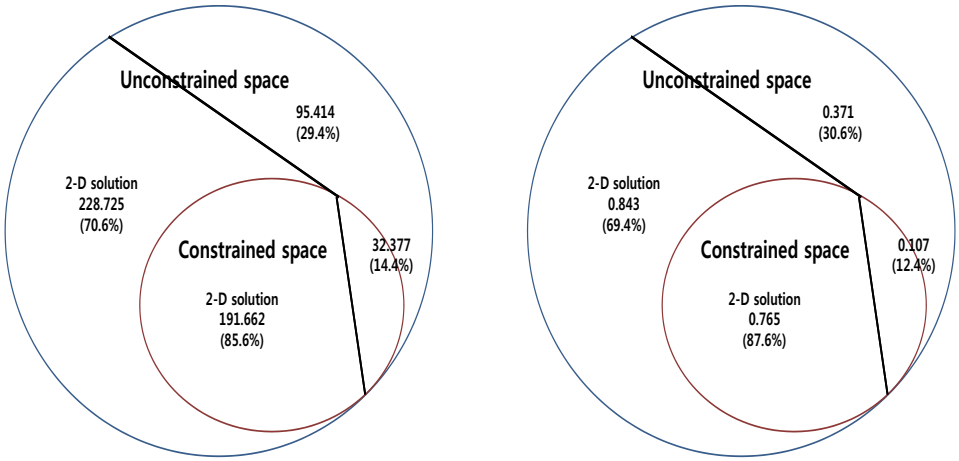


Figure 3.1. 2-D solution in the constrained and the unconstrained space: PCA v.s RDA (Left) and CA v.s CCA (Right) (Greenacre, 2007)

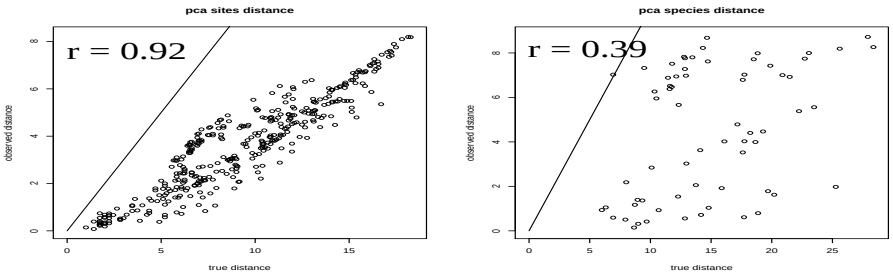
Table 3.1은 네가지 방법의 고유값과 누적비율을 보여주고 있다. 제 2축까지의 누적설명력만 살펴보면, 주성분분석(70.6%), 대응분석(69.4%), 중복분석(85.6%), 정준대응분석(87.6%) 순으로, 중복분석이나 정준대응분석의 설명력이 주성분분석이나 대응분석보다 높다고 할 수 있다. 그러나 이는 Figure 3.1에서 언급한 바와 같이 제한된 공간하에서의 설명력을 나타내기 때문에, 정준대응분석은 대응분석 공간에서, 중복분석은 주성분분석 공간에서, 다시 계산한 수정누적비율로 비교하는 것이 합리적이다.

수정누적설명력은 누적고유값을 중복분석은 주성분분석 전체 고유값의 합 324.139로, 정준대응분석은 대응분석 전체 고유값의 합 1.214로, 나누어서 계산한다. 따라서 2축까지의 수정누적설명력은 주성분분석(70.6%), 대응분석(69.4%), 중복분석(59.2%), 정준대응분석(62.9%)임을 볼 수 있는데, 주성분분석이나 대응분석에 비해서 중복분석이나 정준대응분석의 설명력이 상대적으로 낮음을 볼 수 있다. 그런데, 중복분석이나 정준대응분석의 상대적으로 낮은 설명력은 환경변수 공간상에서 종과 장소를 나타내려는 데서 기인한 것으로 이해할 수 있다. 한편, 이러한 고유값을 이용한 설명력은 모집단에 대한 설명이 아닌 조사된 자료를 이차원 평면에 얼마나 가깝게 나타낼 수 있는지를 나타낼 뿐이다.

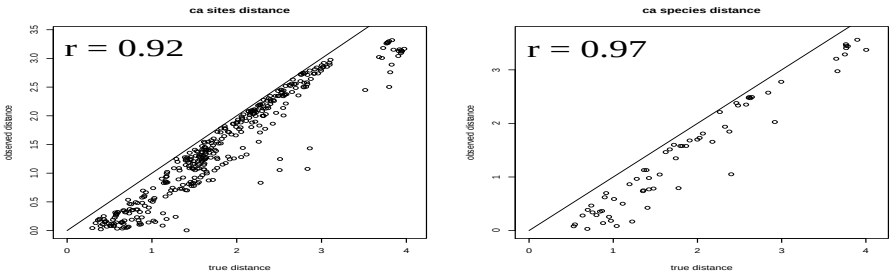
3.2. 거리그래프를 이용한 설명력 비교법

특이값분해를 통해 구한 좌표의 실제거리(true distance)를 가로축에, 차원축소된 공간에서 관찰된 거리(observed distance)를 세로축에 놓고 종 및 장소를 표시했을 때 대각선에 점들이 많이 분포할수록 차원축소된 평면이 조사된 자료를 제대로 표현한다고 할 수 있다. 그런데 시각적 확인이 가능하다는 장점이 있는 반면, 수치적 비교가 어렵다는 한계가 존재한다. 다만 우리는 각 방법들의 상관계수를 비교할 수 있다.

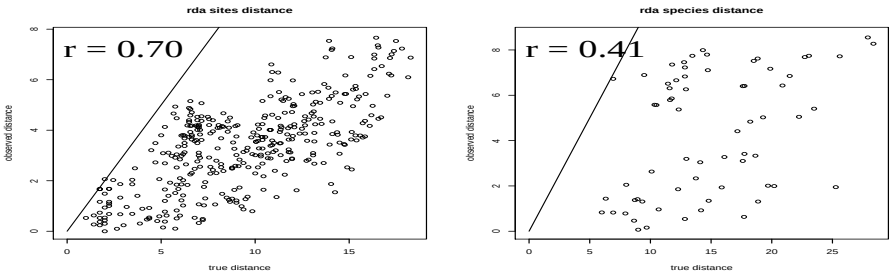
Figure 3.2는 거미 자료에 대한 네가지 서열화방법의 장소와 종들의 거리그래프 및 상관계수를 보여준다. 여기서, 대응분석의 상관계수가 각각 0.92, 0.97로 자료를 이차원 평면에 가장 잘 나타낸다고 볼 수 있다. 그런데, 고유값을 이용한 방법과 같이 모집단 구조가 아닌 자료 (표본) 자체를 2차원 평면에 나타낸 정도만을 알 수 있다는 한계를 지니고 있다.



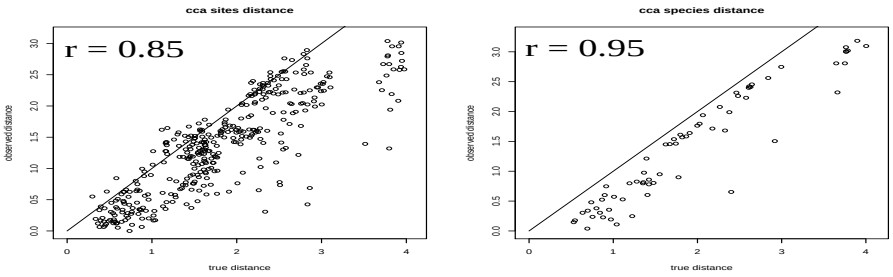
(a) Principal component analysis: sites distance (Left) and species distance (Right)



(b) Correspondence analysis: sites distance (Left) and species distance (Right)



(c) Redundancy analysis: sites distance (Left) and species distance (Right)



(d) Canonical correspondence analysis: sites distance (Left) and species distance (Right)

Figure 3.2. Distance graph for principal component analysis, correspondence analysis, redundancy analysis and canonical correspondence analysis

Table 4.1. Population species emergence frequency

	species 1	species 2	species 3
site 1	10	20	50
site 2	10	50	20
site 3	20	10	50
site 4	20	50	10
site 5	50	10	20

Table 4.2. The distance between sites and species on biplot

	species 1	species 2	species 3
site 1	3	2	1
site 5	2	1	3

4. 출현빈도와 좌표거리를 이용한 서열화방법 비교법

고유값 및 거리그래프를 이용한 설명력 비교방법은 조사된 자료를 이차원 평면상에 근사하게 나타내는 정도를 표현할 뿐이다. 즉, 모집단을 어느 정도 그림에 표현했는지는 알 수가 없어서 서열화방법들을 비교하기에는 한계가 있다. 이에 따라 본 연구에서는 행렬도에 모집단 재현정도를 측정할 수 있는 지표를 소개하고, 모의실험을 통해 서열화방법을 비교한 후, 모의실험에 의해 선택된 방법을 사용하여 거미 자료에 적용하기로 한다.

4.1. 제안 방법

본 연구에서 제안하는 방법의 기본 아이디어는 어떤 종이 특정한 장소들에서 출현빈도가 높다면 이차원 그림 상에서도 다른 장소에 비해 해당 장소들이 서로 가깝게 나타날 것이라는 점을 활용한 것이다. 예를 들어 5개의 장소와 3개의 종으로 구성된 모집단의 출현빈도가 Table 4.1과 같다고 하자. 그리고 Table 4.1에서 2개의 장소를 임의로 선택한 후 서열화방법을 적용하고, 행렬도 좌표를 통해 종과 장소의 거리를 계산한 결과를 Table 4.2라고 하자.

장소1은 모집단에서 종3의 출현빈도가 가장 높고 표본에서도 종3과 거리가 가장 가깝기 때문에 모집단 구조를 잘 재현하고 있다고 하여 장소1은 서열화방법에서 정분류된 곳으로 볼 수 있다. 반면, 장소5의 결과를 살펴보자. 장소5는 모집단에서는 종1의 출현빈도가 가장 높지만 표본에서는 종2와 거리가 가장 가까움을 볼 수 있다. 그러므로 장소5는 서열화방법에서 오분류되었다고 볼 수 있다.

이제, 모의실험에 사용될 자료의 모집단이 가우스 분포(gaussian distribution)를 따른다고 가정하자. 이것은 어떤 종이 최적의 조건에서 가장 많이 출현하고 최적 조건에서 멀어질수록 출현빈도가 떨어짐을 표현하는 가정이라 할 수 있다. 이를 모형으로 나타내면 식 (4.1)과 같다.

$$Y_{ik} = c_k \times \exp\left(-\frac{(X_i - u_k)^2}{2t_k^2}\right), \tag{4.1}$$

여기서 X_i는 i장소에서의 환경변수값, Y_{ik}는 i장소에서 k종의 출현빈도, u_k는 최적값(optimum value), t_k는 허용값(tolerance), c_k는 최대값이다. 환경변수와 종의 수가 많아질수록 식 (4.1)의 모수가 급격히 증가하기 때문에 모형 이해가 어려워진다. 따라서 모형을 단순화하기 위해 Ter Braak (1986)은 다음과 같은 4가지를 가정하였다.

1. 환경변수 변화에 따라 모든 종의 허용값은 동일하다. 즉 $t_k = t$이다.

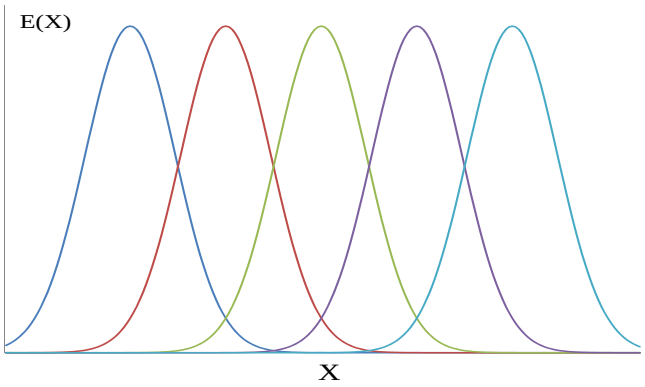


Figure 4.1. Gaussian response curves in the species packing model

2. 환경변수 변화에 따라 모든 종의 가우스 반응(gaussian response)의 최대값은 동일하다. 즉 $c_k = c$이다.
3. 환경변수 변화에 따라 최적값 u_k는 허용값 t_k 보다는 크며 일정한 간격을 유지한다.
4. 환경변수 변화에 따라 조사된 장소는 모집단을 대표한다.

이를 종 패킹 모델(species packing model)이라고 하며 해당 모형의 발현은 식 (4.2)와 같이 나타나며, 이를 그림으로 표현하면 Figure 4.1과 같다.

$$Y_{ik} = c \times \exp\left(-\frac{(X_i - u_k)^2}{2t^2}\right). \tag{4.2}$$

사실, 앞의 4가지 가정은 생태학에서 종의 발현 정도에 대해서는 다소 괴리가 있는 강한 가정이라 할 수 있다. 그런데, Zuur (1999)는 가정 4만 만족하고 나머지 가정이 위배된다고 하더라도, 정준대응분석결과가 강건(robust) 함을 모의실험으로 보인 바 있다. 이러한 결과는, 위의 네가지 가정이 모두 충족되지 않더라도 자료분석을 통해 모집단 구조를 파악할 수 있음을 암시하는 것이다. 하지만, 본 연구에는 단순한 상황하에서 각 방법론의 성능을 비교하기 위해 위의 네 가정을 모두 적용하여 모의실험을 실시하고자 한다.

4.2. 모의실험과 자료분석

4.2.1. 모의실험 모형 단순화를 위해 환경변수에 대한 최적값 u_k를 동일하게 부여하였으며, 환경변수는 2개만을 고려하였다. 모의실험 절차는 다음과 같다.

1. 다중 가우스 회귀(multiple gaussian regression)를 사용하여 2개 환경변수 변화에 따른 3개 종의 반응함수(response function) 값을 100개 발생한다. 즉,

$$Y_{ik} = c \times \exp\left(-\frac{(X_{i1} - \mu_k)}{2t^2} - \frac{(X_{i2} - \mu_k)}{2t^2}\right), \quad i = 1, \ldots, 100,\ k = 1, \ldots, 3$$

   이다.
2. 반응함수 값에 상응하는 다항분포를 발생하여 종-발현 자료 $Y_{100\times3}$ 및 환경자료 $X_{100\times2}$를 생성한다.

Table 4.3. Results of simulation (classification rate, $N = 100$)

methods	$n = 10$	$n = 20$	$n = 50$
principal component analysis	0.696	0.723	0.740
correspondence analysis	0.912	0.936	0.948
redundancy analysis	0.711	0.711	0.711
canonical correspondence analysis	0.808	0.824	0.894

Table 4.4. Results of spider data analysis (classification rate, $N = 28$)

methods	$n = 5$	$n = 10$	$n = 15$
principal component analysis	0.740	0.751	0.753
correspondence analysis	0.818	0.842	0.855
redundancy analysis	0.712	0.711	0.715
canonical correspondence analysis	0.776	0.785	0.791

3. 종별로 출현빈도가 상대적으로 가장 높은 장소를 찾는다.
4. 생성된 종-발현 자료 $Y_{100\times3}$ 및 환경자료 $X_{100\times2}$를 모집단으로 놓고, 표본을 10개, 20개, 50개를 뽑아 4가지 서열화방법(주성분분석, 대응분석, 중복분석, 정준대응분석)을 사용하여 종과 장소의 좌표를 계산한다.
5. 단계 4에서 계산된 좌표를 이용하여 종별로 상대적으로 가장 가까운 장소를 찾는다.
6. 3과 5를 비교하여 정분류율을 계산한다.
7. 단계 4부터 단계 6까지를 100회 반복하여 정분류율의 평균을 계산한다.

모의실험을 통해 주요 서열화방법의 정분류율을 계산한 결과는 Table 4.3과 같다. 여기서 정분류율은 표본에서 나타난 종과 장소의 관계가 모집단과 동일한 빈도의 평균값이다(반복 = 100). Table 4.3의 결과, 모든 서열화방법들에서 표본크기가 커짐에 따라 모집단 설명력이 커짐을 볼 수 있다. 한편, 대응분석이 정분류율이 가장 높고, 다음으로 정준대응분석이 높음을 볼 수 있는데, 표본이 커짐에 따라 두 방법의 격차가 줄어들고 있음을 알 수 있다. 그런데, 정준대응분석이 환경변수를 같은 공간상에 표현해서 종과 장소와의 관계를 보다 직관적으로 알 수 있게 많은 정보를 제공한다는 점에서, 약간의 정분류율을 포기하더라도 정준대응분석을 사용한 서열화 방법이 대응분석을 사용한 서열화 방법보다는 보다 더 현실적으로 유용하다고 생각된다.

4.2.2. 자료분석 모의실험 결과를 기초하여 거미자료에 대한 자료분석을 살펴보자. 서열화 결과를 비교하기 위해 앞의 거미자료를 모집단으로 가정하고 표본을 추출하여 정분류율을 계산하기로 한다. 다만, 서열화방법간 차이를 확실히 보기위해 장소별 출현빈도가 높은 3개 종(trochosa terricola, pardosa monticola, alopecosa accentuata), 종 출현에 주로 영향을 3개 환경변수(이끼 분포 비율, 빛 반사 비율, 엽채류 분포 비율)만을 고려하였다.

Table 4.4는 모의실험 절차에 따라 100회 반복하여 정분류율의 평균값을 계산한 것이다. 모의실험 결과와 유사하게 표본크기가 커짐에 따라 모집단 설명력이 커지고 있으며, 대응분석과 정준대응분석이 주성분분석과 중복분석에 비해 정분류율이 높음을 볼 수 있다. Figure 4.2는 모의실험에 의해 모집단 재현성이 높은 것으로 평가된 대응분석과 정준대응분석, 재현성이 낮은 것으로 평가된 주성분분석과 중복분석을 사용한 거미 자료의 행렬도이다. Figure 4.2를 보면 대응분석계열(대응분석, 정준대응분석)과 주성분분석계열(주성분분석 중복분석)의 행렬도가 확연히 다르고 계열내에서는 유사함을 볼 수 있다. 어느

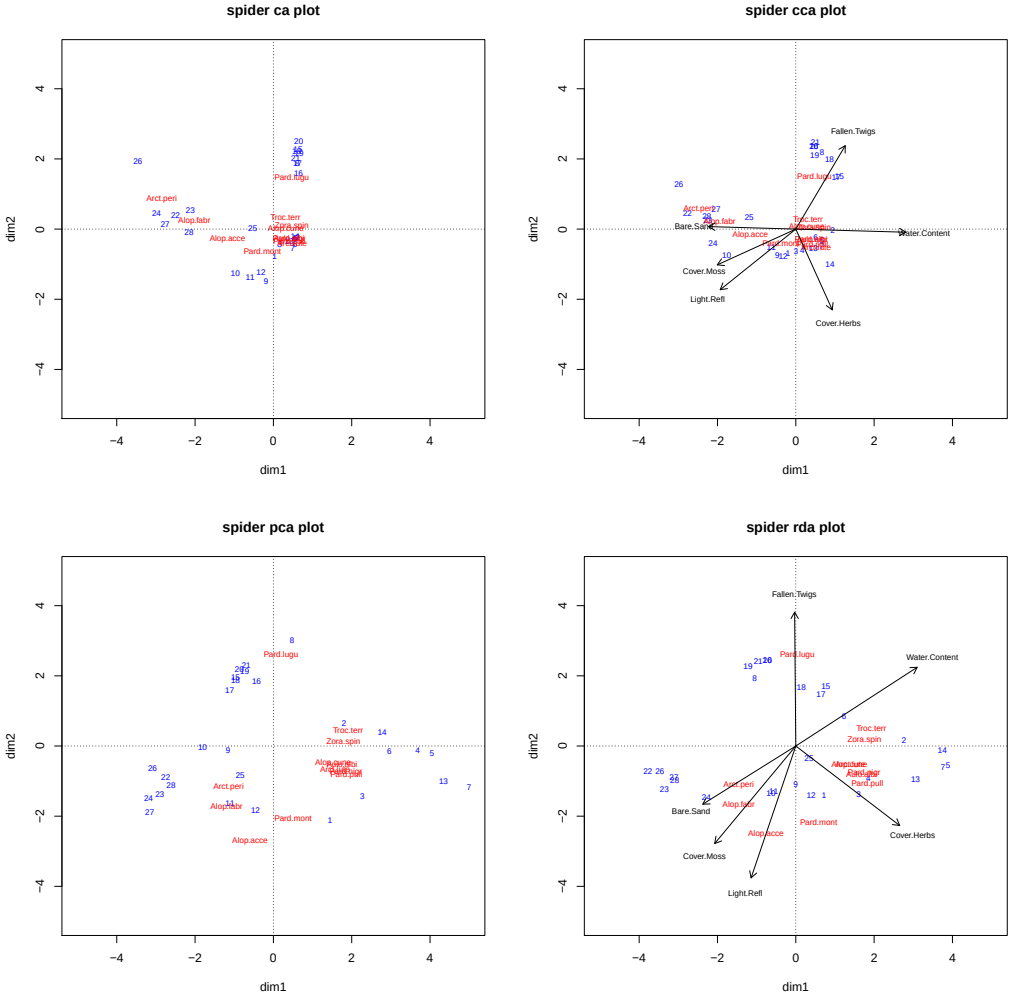


Figure 4.2. Upper panel:Biplot of correspondence analysis (Left) and tri-plot of canonical correspondence analysis (Right), Lower panel:Biplot of principal component analysis (Left) and tri-plot of redundancy analysis analysis (Right)

쪽이 모집단을 잘 재현하고 있는지는 행렬도만으로는 확인할 길이 없지만, 앞의 모의실험 결과에 비추어 볼 때 대응분석계열이 모집단의 특성을 잘 반영하고 있다고 판단할 수 있다. 다만 대응분석계열내에서 대응분석은 가로축과 세로축의 환경변수적 의미를 찾아야 하는 이유로 환경에 따른 종과 장소의 관계를 직접적으로 규명하기 곤란하지만, 정준대응분석은 환경변수를 같은 행렬도에 나타내 주기 때문에 환경변화에 따른 종과 장소의 변화를 시각적으로 볼 수 있는 장점을 지닌 분석임을 알 수 있다.

5. 결론

서열화방법은 종과 환경의 관계를 시각적으로 보기 위한 방법의 일종으로, 많은 연구자들이 지속적으로

종-발현에 대한 여러 효율적인 분석을 제안하여 왔다. 그런데, 종의 출현빈도는 환경변수가 변함에 따라 계속 증가 하거나 감소하는 경향을 보이지 않고 최적값을 중심으로 분포하는 경향을 보인다. 그러므로 대응분석이나 정준대응분석으로 서열화 방법을 제시하는 것이 모집단에 근사하게 종과 환경의 관계를 나타낼 수 있다고 하였다 (Legendre와 Legendre, 2012). 그러나 대부분의 연구가 방법의 우수성에 대한 객관적인 지표를 제시하지 못하고 주관적인 판단에 그치는 경우가 많았다. 일부에서는 고유값과 거리그래프 등을 이용하여 서열화방법 비교를 시도 하였지만, 이것도 또한 모집단이 아닌 표본을 2차원 평면에 나타내는 정도만을 알 수 있다는 한계가 있었다.

본 연구에서는 모집단에서 어떤 종이 특정한 장소에서 출현빈도가 높다면 행렬도에서 가깝게 나타날 것이라는 가정하에 정분류율을 계산하였고, 자료를 가우스분포에서 발생하여 여러 서열화방법에 적용하여 모의실험을 실시하였다. 이를 통해 가우스분포를 따르는 종의 출현빈도는 주성분분석이나 중복분석 보다는 대응분석이나 정준대응분석을 통한 서열화방법이 모집단에 근사하게 종과 장소의 관계를 나타낼 수 있음을 살펴볼 수 있었다. 특히, 정준대응분석은 환경정보를 추가적으로 더 활용할 수 있고 환경변수를 동일한 평면에 나타낼 수 있어서, 대응분석 보다 다소 낮은 정분류율에도 불구하구 종(개체)의 발현에 대해 보다 현실적 통찰력을 제공하는 분석임을 확인하였다.

References

Benzecri, J. P. (1973). *L'Analyse des Donnees, Volume II, L'Analyse des Correspondances*, Paris, France: Dunod.

Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin, *Ecological Monographs*, **27**, 325–349.

Greenacre, M. (2007). *Correspondence Analysis in Practice*, Chapman & Hall, London.

Hill, M. O. (1973). Reciprocal averaging: an eigenvector method of ordination,*Journal of Ecology*, **61**, 237–249.

Jackson, D. A. and Somers, K. M. (1991). Putting things in order: The UPs and Downs of detrended correspondence analysis, *The American Naturalist*, **137**, 704–712.

Jeong, H. C. (2012). A study of canonical correspondence analysis for community ordination, *Journal of the Korean Data Analysis Society,* **14**, 2385–2395. (in Korean)

Kim, D. and Jeong, H. C. (2013). On the application of reciprocal averaging in correspondence Analysis, *Journal of the Korean Data Analysis Society*, **15**, 3087–3099. (in Korean)

Legendre, P. and Legendre, L. (2012). *Numerical Ecology*, Elsevier.

Parmer, M. W. (1993). Putting things in even better order: The advantages of canonical correspondence analysis, *Ecology*, **74**, 2215–2230

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

Zuur, A. F. (1999). *Dimension reduction techniques in community ecology with applications to spatio-temporal marine ecological data*, PhD thesis, Aberdeen, Scotland: University of Aberdeen.

생태학의 통계적 서열화 방법 비교에 관한 연구

고현석[a] · 전명식[b] · 정형철[c,1]

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

(2014년 10월 22일 접수, 2014년 12월 23일 수정, 2015년 1월 20일 채택)

요 약

군생태학에서 종, 장소 그리고 환경변수의 관계를 시각적으로 보기 위해 대응분석, 정준대응분석 등 다양한 서열화방법들을 사용한다. Ter Braak (1986), Jackson 등 (1991), Parmer (1993) 등은 고유값 및 거리그래프를 이용하여 서열화방법들을 비교하고 있는데, 이 방법들은 조사된 데이터에 근거하고 있기 때문에, 모집단과 행렬도의 관계를 보여주지는 못한다. 따라서, 본 논문에서는 행렬도에 모집단 정보의 표현정도를 측정하는 방법을 소개하고, 이를 활용하여 서열화방법들을 객관적으로 비교하였다. 비교결과, 정준대응분석은 대응분석과 유사한 정분류율을 유지하면서도 환경정보를 이차원 공간에 표현할 수 있는 장점을 지닌 분석임을 확인하였다.

주요용어: 거리그래프, 주성분분석, 대응분석, 중복분석, 정준대응분석, 서열화방법, 서열분석, 정준서열분석, 특이값분해.

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