The Korean Journal of Applied Statistics
2021, Vol. 34, No. 6, 945–956

DOI: https://doi.org/10.5351/KJAS.2021.34.6.945
Print ISSN1225-066X / Online ISSN 2383-5818

Equivalence study of canonical correspondence analysis by weighted principal component analysis and canonical correspondence analysis by Gaussian response model

Hyeong Chul Jeong[a] [1],

[a]Division of Data Science, University of Suwon

Abstract

In this study, we considered the algorithm of Legendre and Legendre (2012), which derives canonical correspondence analysis from weighted principal component analysis. And, it was proved that the canonical correspondence analysis based on the weighted principal component analysis is exactly the same as Ter Braak's (1986) canonical correspondence analysis based on the Gaussian response model. Ter Braak (1986)'s canonical correspondence analysis derived from a Gaussian response curve that can explain the abundance of species in ecology well uses the basic assumption of the species packing model and then conducts generalized linear model and canonical correlation analysis. It is derived by way of binding. However, the algorithm of Legendre and Legendre (2012) is calculated in a method quite similar to Benzecri's correspondence analysis without such assumptions. Therefore, if canonical correspondence analysis based on weighted principal component analysis is used, it is possible to have some flexibility in using the results. In conclusion, this study shows that the two methods starting from different models have the same site scores, species scores, and species-environment correlations.

Keywords: weighted principal component analysis, Gaussian response curve, canonical correspondence analysis, singular value decomposition, linear combination score

1. 서론

정준대응분석(canonical correspondence analysis, CCA)은 Ter Braak (1986)에 의해 최초로 제안되었다. Ter Braak (1986)은 잠재변수를 사용한 제한적 가우시안 반응곡선 모형을 사용하여, 방향성 축(gradient) 별로 잠재변수의 반응값(site score)과 여러 관찰된 생물 종들의 최대 발현량(species score)을 추정하였다. 그리고, 장소와 종, 환경을 모두 2차원 공간에 표현하여 이들의 대응 관계를 살펴보았으며, 이를 정준대응분석이라 하였다(Jeong, 2012). 또한, 가우시안 반응곡선에 기반 한 정준대응분석 패키지 CANOCO를 발표하였다 (Ter Braak, 1988). 앞의 가우시안 반응곡선의 특징에 대해서는 Jeong (2018b)를 참고할 수 있다. Hegde와 Nail (1999, 2008)은 Ter Braak (1988)의 CANOCO를 SAS에서 구동 시킬 수 있는 SAS/IML 코드를 발표하였으며, Aart와 Smeenk (1975) 자료를 사용하여 SAS/IML에서의 계산과정을 친절하게 보인 바 있다 (Hegde, 2012).

그런데 Ter Braak (1986)의 알고리즘은 가우시안 반응곡선회귀 추정에 대한 일반화선형모형과 species packing model을 가정하며, 그 유도과정이 상당히 난해하다. 그리고 최종 결과는 정준대응분석의 기초행렬

This paper was supported by the research grant of the University of Suwon in 2020.
[1] Division of Data Science, University of Suwon, 17,Wauan-gil, Bongdam-eup, Hwaseong-si, Gyeonggi-do 18323, Korea. E-mail: jeong.hc@suwon.ac.kr

Published 31 December 2021 / journal homepage: http://kjas.or.kr

Table 1: Structure of data for canonical correspondence analysis

Species abundance table	Environmental variable	Gradient (latent variable)
$Y_{n\times p}=[y_{ik}]$	$X_{n\times q}=[x_{ij}]$	$Z_{n\times m}=[z_{il}]$

(fundamental matrix of CCA)을 구성하는 분할행렬 W에 대한 정준상관분석(canonical correlation analysis)으로 마무리된다 (Jeong, 2020). 한편, Huh (1998, 2011)는 Hayashi의 수량화 3법을 설명하면서 범주형 분할표에 대한 정준상관분석이 Benzecri (1973)의 대응분석과 동일함을 보인바 있다 (Kim과 Jeong, 2013). 즉, 대응분석과 정준상관분석은 서로 밀접한 관련성을 지니고 있기 때문에, 정준상관분석을 사용하는 Ter Braak (1986)의 방법이 설명변수가 있는 정준상관분석이란 의미로 정준대응분석이라 칭하는 것에 이견이 없다. 하지만 Ter Braak (1986)의 CCA가 Benzecri (1973)의 대응분석과 수리적으로 어떤 관련성을 지니고 있는지에 대해서는 약간의 모호성과 난해함이 있다고 할 수 있다.

한편, Legendre와 Legendre (2012)는 Ter Braak (1986)의 가우시안 반응곡선과는 전혀 다른 방법으로 CCA를 그들의 저서 *Numerical Ecology*에 간략히 언급하였다 (Ko 등, 2015a, 2015b). 여기서는 종 발현 자료를 종속변수로 고려하고, 환경변수로 다변량회귀분석(multivariate regression) 한 후 차원 축소된 공간 자료에 대한 가중주성분분석을 실시한다. 이는 회귀분석으로 차원축소 된 자료에 대해 Benzecri (1973)의 대응분석 방법을 적용하는것과 동일하다. Legendre와 Legendre (2012)알고리즘의 장점은 복잡한 가우시안 반응곡선과 일반화선형모형을 사용하지 않으면서도 Ter Braak (1986)의 CCA와 동일한 결과를 유도한다는 점이다. 그리고, 대응분석과 동일하게, 표준좌표, 주좌표, 대응좌표 등의 개념을 상황에 따라 적용할 수 있다는 장점을 가지고 있다.

정준대응분석을 실시할 수 있는 프로그램 역시 다양한데, CANOCO 이외에 R에서 ANACOR (2009), VEGAN (2011), ADE4 (2012) 등을 수행할 수 있다 (Jeong 2020). 각각의 프로그램이 어떤 알고리즘에 기초하여 실행되는가는 명확하진 않지만, 스케일과 고유치 반영 정도, 거리행렬에 따른 다양한 옵션 차이가 있을 뿐 그 결과는 별반 차이가 없는 편이다.

본 연구에서는 Legendre와 Legendre (2012)의 알고리즘을 살펴보고, Ter Braak (1986)의 CCA와 수리적 동일함을 보이는 문제를 다루었다. 2장에서는 Ter Braak (1986)의 정준대응분석과 Legendre와 Legendre (2012)의 정준대응분석을 소개하였다. 특히, 가중주성분분석의 정준대응분석을 위해 Benzecri (1973)의 대응분석을 간략히 언급하였다. 3장에서는 기초행렬의 대칭행렬에 대한 고유치 분해를 통해 두 모형의 동일성을 다루었으며, 4장에서는 간략한 수치적 예를 소개하였다. 본 연구의 이하 전개과정에서는 가우시안반응모형에 기초한 Ter Braak (1986)의 정준대응분석(Ter Braak's CCA)을 TCCA로, 가중주성분분석에 기초한 Legendre와 Legendre (2012)의 정준대응분석(Legendre's CCA)를 LCCA라 표기하기로 한다.

2. 정준대응분석 모형

2.1. 자료구조 및 수식

생태학에서 다루는 species-abundance-environmental 자료는 Table 1과 같다.

$X_{n\times q}$의 x_{ij} $(i=1,\ldots,n;\ j=1,\ldots,q)$는 i 장소(site)의 j번째 환경변수에 대한 측정(관찰) 값, $Y_{n\times p}$의 y_{ik} $(i=1,\ldots,n;\ k=1,\ldots,p)$는 i 장소에서 발견된 k번째 종(species)의 출현(관찰) 빈도, $Z_{n\times m}$의 z_{il} $(i=1,\ldots,n;\ l=1,\ldots,m)$,은 i 장소에서 추정된 l 번째 축의 방향(gradient)값이다. 여기서, X와 Y는 측정(관찰)된 값이지만, Z는 미지(unknown)의 값으로 추정해야 할 부분이다 (Jeong, 2020). 즉, 종-발현-환경 자료는 종속변수 역할을 하는 분할표 Y, 독립변수 역할을 하는 환경값 X, 환경값 X의 선형결합에 의해 차원축소되어 나타나는 잠재변수 Z로 구성된 자료이다. 이러한 데이터는 장소별로 여러 생물 종 들의 출현 빈도가 나타난 분할표 Y에

연속형 혹은 이산형의 환경 관찰값이 추가되어, 생태학 상 환경 변수 X에 따라 종 발현 정도 Y가 어떤 영향을 받는가를 밝히고자 하는 연구를 위해 도출된 것이다.

이제, CCA를 위해 다음의 수식을 사용하기로 한다.

$Y = [y_{ik}]$ 는 $i = 1, \ldots, n;\ k = 1, \ldots, p$ 로 i 장소에서 k 종의 출현빈도 자료

$\underline{y}_{i+}=\{y_{i+}\}$는 i 장소별 모든 종의 행(장소) 합 벡터

$\underline{y}_{+k}=\{y_{+k}\}$는 k 종별 모든 장소에서의 열(종) 합 벡터

$N = y_{++}$ 는 모든 종의 발현(관찰) 총 합

$P = [p_{ik}] = Y/N$ 는 상대도수 행렬

$D_r^* = \text{diag}\,(y_{1+}, y_{2+}, \ldots, y_{n+})$ 행 합 대각행렬

$D_r = D_r^*/N$ 행 비율 대각행렬

$\underline{r}=\{p_{i+}\}=\underline{y}_{i+}/N$ 행 비율 벡터 혹은 행 가중치 벡터(열 프로파일의 중심)

$D_c^* = \text{diag}\,(y_{+1}, y_{+2}, \ldots, y_{+p})$ 열 합 대각행렬

$D_c = D_c^*/N$ 열 비율 대각행렬

$\underline{c}=\{p_{+k}\}=\underline{y}_{+k}/N$ 열 비율 벡터 혹은 열 가중치 벡터(행 프로파일의 중심)

$\underline{a}_i = (y_{i1}, \ldots, y_{ip})'/y_{i+}$ 로 행(site) 프로파일 벡터

$\mathbf{B} = [\underline{b}_i] = \{\underline{a}_i - \underline{c}\} = D_r^{-1}(P - \underline{r}\,\underline{c}')$ 는 행프로파일과 중심 프로파일과의 차이 행렬

$X = [x_{ij}]$ 는 $i = 1, \ldots, n;\ j = 1, \ldots, q$ 로 i 장소에서 q개의 환경변수에 대한 환경 관찰값 자료

$X_c = X - 1_n D_r^* X$ 환경 자료에 대한 가중 중심화 행렬, 여기서 $1_n = [1]_{n\times n}$ 행렬

$X_s = X_c \text{Diag}\,(V)^{-1/2}$ 환경자료에 대한 가중 표준화 행렬, 여기서 $V = X_c D_r^* X_c$는 가중분산

$Z = [z_{il}]$은 $i = 1, \ldots, n;\ l = 1, \ldots, m$ 로 i 장소에서 l 번째 방향 축의 잠재값

여기서, 특정 l번째 방향 축에 대해,

$\underline{u} = \{u_k\}$는 $k = 1, \ldots, p$로 종들의 최적의 환경값 벡터

$\underline{\beta} = \{\beta_j\}$는 $j = 1, \ldots, q$로 환경변수에 대한 정준계수(회귀계수) 벡터

$\underline{z} = \{z_i\} = X_s\underline{\beta}$ 는 환경변수에 대한 선형결합(linear combination) 벡터

2.2. Ter Braak's canonical correspondence analysis

Ter Braak's CCA 유도과정은 Jeong (2020)에서 자세히 언급되었으며, 여기서는 간략히 그 결과를 소개하기로 한다. Ter Braak's CCA는 q개의 표준화된 환경변수 $x_j, j = 1, \ldots, q$의 선형결합으로 만들어진 잠재변수 $z_l, l = 1, \ldots, m$ 들에 대한 제한적 가우시안 모형에서 출발한다.

다음 식 (2.1)은 $l = 1$(첫번째 축)에 대한 제한적 가우시안 모형이다.

$$E(y_{ik}) = \mu_{ik} = c\ \exp\left\{-\frac{(z_i - u_k)^2}{2t^2}\right\} = \exp\left(\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2\right), \tag{2.1}$$

$$z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} = \sum_{j=1}^{q} \beta_j x_{ij} = X_s\underline{\beta}.$$

이제, 종 발현 y_{ij}가 포아송분포를 따른다고 놓고, 우도함수 $L(y;\mu_{ik}) = \Pi_k\Pi_i e^{\mu_{ik}}\mu_{ik}^{y_{ik}}/y_{ik}!$에 대해 $\partial logL/\partial u_k = 0$, $\partial logL/\partial\beta_j = 0$을 계산하여 다음의 관계식으로 $\underline{\beta}$, u_k, z_i를 유도할 수 있다 (Jeong, 2020).

$$(A)\ \lambda u_k = \sum_i^n \frac{y_{ik}}{y_{+k}} z_i\ \ (B)\ z_i^* = \sum_k^q \frac{y_{ik}}{y_{i+}} u_k\ \ (C)\ \underline{\beta} = (X_s' D_r^* X_s)^{-1} X_s' D_r^* z^*\ \ (D)\ Z = X_s\underline{\beta}$$

여기서 m개의 축 별로($l = 1, \ldots, m$), u_k를 종 점수(species score), z_i를 장소점수(site score)라 하며, 잠재변수인 z_i를 환경 변수들의 선형결합으로 계산된다는 의미로 선형결합점수(linear combination score)라 칭하기도 한다. 위의 (A)~(D) 관계식에 대해 반복연산으로 각 score를 계산할 수 있지만, 다음의 분할행렬 W 에 대한 고유치 분해로 한꺼번에 여러 잠재축에 대한 종 점수, 장소점수 그리고 설명력 등을 계산할 수 있다.

$$W = \begin{array}{c|c} S_{11} = D_c^* & S_{12} = Y'X_s \\ \hline S_{21} = X_s'Y & S_{22} = X_s'D_r^*X_s \end{array}$$

즉, W 행렬에 대해 정준상관분석을 실시하여 여러 잠재변수 차원에 대한 score를 유도하는 것이 Ter Braak (1986)의 CCA이다. 특히 $A = D_c^{*-1/2}(Y'X_s)(X_s'D_r^*X_s)^{-1/2}$로 놓으면, A에 대한 특이치 분해(singular value decomposition, SVD)

$$D_c^{*-1/2}(Y'X_s)(X_s'D_r^*X_s)^{-1/2} = UD_{\sqrt{\lambda}}V'$$

에 따라 $u = D_c^{-1/2}U, \beta = (X_s'D_rX_s)^{-1/2}V, Z = X_s\beta$가 계산된다. 여기서 $\beta = [\underline{\beta}_l]$로 m개의 잠재변수 축에 따라 추정된 벡터 $\underline{\beta}$를 모아놓은 행렬을 의미하며, $A = D_c^{*-1/2}(Y'X_s)(X_s'D_r^*X_s)^{-1/2}$를 Ter Braak's CCA의 기초행렬(fundamental matrix of TCCA)이라 한다 (Hegde와 Naik, 1999; Jeong, 2020).

한편, $D_c = N^{-1}D_c^*$ 라는 사실에 의해,

$$A = N^{-1/2}D_c^{-1/2}(Y'X_s)(X_s'ND_rX_s)^{-1/2} = D_c^{-1/2}(P'X_s)(X_sD_rX_s)^{-1/2}$$

이다. 이제, Legendre와 Legendre (2012)의 알고리즘과 비교를 위해 A 대신 전치행렬 A'을 T_b 행렬이라 놓고 T_b에 대한 SVD를 생각하자.

$$T_b = (X_s'D_rX_s)^{-1/2}(X_s'P)D_c^{-1/2} = U_tD_{\sqrt{\lambda}}V_t' \tag{2.2}$$

기초행렬 T_b에 의한 Ter Braak's CCA 최종 결과는 다음과 같다.

- Species Score : $u = D_c^{-1/2}V_tD_{\sqrt{\lambda}}$ (주좌표 개념)
- Site Score : $Z = X_s(X_s'D_rX_s)^{-1/2}U_t$
- Species-environment correlation : Cor $(D_r^{1/2}X_s, D_r^{1/2}Z) = X_s'D_rZ$

본 연구에서는 Legendre's CCA와 비교를 위해 T_b를 Ter Braak's CCA의 기초행렬(fundamental matrix)이라 칭하기로 한다.

2.3. Legendre's canonical correspondence analysis

Benzecri의 대응분석은 분할표에 대한 가중 주성분분석으로 알려져 있으며, Hill (1973)은 Benzecri의 재형성식을 사용하여 상호평균법(reciprocal averaging)을 제안한 바 있다 (Kim과 Jeong, 2013). 본 절에서는 Benzecri의 대응분석 알고리즘에 충실한 Legendre's CCA를 소개하기로 한다

$Y = [y_{ik}]$ 에 대한 Pearson 카이제곱 통계량은 $q_{ik} = (p_{ik} - p_{i+}p_{+k})/\sqrt{p_{i+}p_{+k}}$ 를 기초로 하며,

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{n}\sum_{k=1}^{p}\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \\ &= \sum\sum\frac{\left(\frac{y_{ik}-y_{i+}y_{+k}}{N}\right)^2}{\frac{y_{i+}y_{+k}}{N}} = N\sum\sum\left(\frac{p_{ik}-p_{i+}p_{+k}}{\sqrt{p_{i+}p_{+k}}}\right)^2 = N\sum\sum q_{ik}^2 \end{aligned}$$

이다. 여기서, $Q = [q_{ik}]$는

$$Q = [q_{ik}] = \frac{p_{ik} - p_{i+}p_{+k}}{\sqrt{p_{i+}p_{+k}}} = \frac{(p_{ik} - p_{i+}p_{+k})/p_{i+}}{\sqrt{p_{i+}p_{+k}}/p_{i+}} = \frac{p_{k|i} - p_{+k}}{\sqrt{p_{+k}/p_{i+}}} = \sqrt{p_{i+}}\left[\frac{p_{k|i} - p_{+k}}{\sqrt{p_{+k}}}\right]$$
$$= D_r^{-1/2}(P - \underline{r}\,\underline{c}')D_c^{-1/2} \quad \text{where} \quad p_{k|i} = p_{ik}/p_{i+}$$

라는 행렬로 표현할 수 있다.
이제, Q^*를 다음과 같이 놓자.

$$Q^* = [q_{ik}^*] = \left[\frac{p_{k|i} - p_{+k}}{\sqrt{p_{+k}}}\right] = \frac{1}{\sqrt{p_{i+}}}\sqrt{p_{i+}}\left[\frac{p_{k|i} - p_{+k}}{\sqrt{p_{+k}}}\right]$$
$$= D_r^{-1/2}Q = D_r^{-1}(P - \underline{r}\,\underline{c}')D_c^{-1/2}$$

행 프로파일(site profile) $\underline{a}_i$ 벡터가 주어지면, $\underline{r} = \{p_{i+}\}, i = 1, \ldots, n$ 벡터는 행의 가중치(site mass)가 되며, n개 행(site) 프로파일의 중심은 $\underline{c} = \sum_{i=1}^{n} r_i a_i = \{c_k\}, k = 1, \ldots, p$ 벡터로 열의 가중치(species mass)가 되어, 위의 χ^2는 다음과 동일한 의미를 지닌다.

$$N \sum_i r_i \sum_k \frac{(a_{ik} - c_k)^2}{c_k} = N \times \sum_i (\text{ 행가중치 })(\text{ 행프로파일의 중심과의 카이제곱거리 })^2$$

이에 따라, 행 프로파일이 중심 프로파일과 떨어진 (절대)거리 행렬 **B**는 앞의 수식에서 언급한 바와 같이,

$$\mathbf{B} = [\underline{b}_i] = \underline{a}_i - \underline{c} = p_{ik}/p_{i+} - p_{+k} = \left\{p_{i+}^{-1}\right\}\{p_{ik} - p_{i+}p_{+k}\} = D_r^{-1}(P - \underline{r}\,\underline{c}')$$

로 표현할 수 있다.

2.3.1. Benzecri's correspondence analysis

Benzecri's 대응분석은 행 프로파일 $\underline{a}_i$가 중심으로부터 떨어진 정도를 나타내는 $\underline{b}_i,\ i = 1, \ldots, n$ 벡터들을 n개 (행의 개수)의 데이터로 간주하고, 이들을 가중유클리디안 공간에서 $v'D_c^{-1}v = 1, 1_p'v = 0$의 성질을 만족하는 단위 벡터 v에 사영(projection)하여 제곱사영값을 최대로 하는 단위벡터 v와 그 사영값을 찾는 방법이다 이를 위해 데이터 **B**를 열 표준화를 하여 $\mathbf{B}D_c^{-1/2}$로 변경시키고, 열 비율 벡터로 표준화된 $\mathbf{B}D_c^{-1/2}$을 $D_c^{-1/2}v$ 축에 사영한다 (Huh, 2011; Kim과 Jeong, 2013). 이러한 대응분석을 행가중치와 열 축 표준화를 고려한다는 점에서 전통적 주성분분석과 비교하여 가중주성분분석이라 한다. 결국, 대응분석에서,

$$Q^* = \mathbf{B}D_c^{-1/2} = D_r^{-1}(P - \underline{r}\,\underline{c}')D_c^{-1/2} = D_r^{-1/2}D_r^{-1/2}(P - \underline{r}\,\underline{c}')D_c^{-1/2} = D_r^{-1/2}Q \tag{2.3}$$

는 매우 중요한 역할을 하며, 이를 대응분석에서 다루어야 할 표준화 데이터로 취급한다.

이제, $B_z = D_r^{1/2}\mathbf{B}D_c^{-1/2}$, $v_z = D_c^{-1/2}v$ 로 놓으면, $B_z(= Q)$의 특이치분해(SVD)는,

$$B_z = D_r^{1/2}\mathbf{B}D_c^{-1/2} = D_r^{-1/2}(P - \underline{r}\,\underline{c}')D_c^{-1/2} = U_z D_{\sqrt{\lambda}} V_z' \tag{2.4}$$

이며, 행(site) 사영은 $\mathbf{B}D_c^{-1/2}$를 $D_c^{-1/2}v$, 열(species) 사영은 열좌표 $a_j^* = (0, \ldots, 1, \ldots, 0)$를 중심화한 벡터 $(I_p - I_p\underline{c}')D_c^{-1/2}$ 를 $D_c^{-1/2}v$에 사영한 것으로 식 (2.4)를 사용하면 다음과 같다.

$$\text{행 사영} : \mathbf{B}D_c^{-1}v = D_r^{-1/2}U_z D_{\sqrt{\lambda}}, \quad \text{열 사영} : (I_p - I_p\underline{c}')D_c^{-1}v = D_c^{-1}v = D_c^{-1/2}V_z.$$

그런데 CCA 결과와 비교를 위해, 열 프로파일의 사영결과를 주좌표(principal coordination), 행 좌표의 사영을 표준좌표(standardize coordination)로 간주하여, Table 1의 Y 자료에 대한 사영결과를 아래와 같이 나타내기로 한다 (Ko 등, 2015a).

- Species Score(열 사영) : $D_c^{-1/2}V_z D_{\sqrt{\lambda}}$ (주좌표 개념)
- Site Score(행 사영) : $D_r^{-1/2}U_z$

2.3.2. Legendre's canonical correspondence analysis

환경변수 $X = \{x_{ij}\}$를 고려한 Legendre's CCA를 살펴보기로 한다. 우선, 환경변수 X_s에 대한 다음의 다변량회귀모형을 생각한다.

$$Q^* = X_s\beta + \mathrm{E} \tag{2.5}$$

여기서, Q^*는 상대돗수와 독립모형과의 차이인 $(P - \underline{r}\,\underline{c}')$에 가중치 D_r^{-1}이 부여된 형태로 식 (2.3)의 다양한 모습을 취하며, 위의 식 (2.5)에서 $\hat{\beta} = (X_s'D_rX_s)^{-1}X_s'D_rQ^*$ 로 가중회귀분석에 의해 추정된다. 여기서 $\hat{\beta}$는 $q \times p$ 행렬이며, $\hat{\beta} = (\underline{\hat{\beta}}_1, \underline{\hat{\beta}}_2, \cdots \underline{\hat{\beta}}_p)$임을 알 수 있다. 가중회귀분석에 의해,

$$\hat{Q}^* = X_s(X_s'D_rX_s)^{-1}X_s'D_rQ^* = X_s\hat{\beta} \tag{2.6}$$

이며, 식 (2.3)의 $Q^* = D_r^{-1/2}Q$, $Q = D_r^{1/2}Q^*$로 부터, $\hat{Q}^* = D_r^{-1/2}\hat{Q}$, $\hat{Q} = D_r^{1/2}\hat{Q}^*$ 로 놓고, 식 (2.6)으로 부터,

$$\begin{aligned} D_r^{-1/2}\hat{Q} &= X_s(X_s'D_rX_s)^{-1}X_s'D_rD_r^{-1/2}Q \\ \hat{Q} &= D_r^{1/2}X_s\hat{\beta} = D_r^{1/2}X_s(X_s'D_rX_s)^{-1}X_s'D_r^{1/2}Q \\ &= D_r^{1/2}X_s(X_s'D_rX_s)^{-1}X_s'(P - \underline{r}\,\underline{c}')D_c^{-1/2} \\ &= D_r^{1/2}\hat{B}D_c^{-1/2} \quad \text{where} \quad \hat{B} = X_s(X_s'D_rX_s)^{-1}X_s'(P - \underline{r}\,\underline{c}') \end{aligned} \tag{2.7}$$

이다. 이제, $\hat{Q}$의 의미를 살펴보자. $\hat{Q}$는 $\hat{Q} = D_r^{1/2}\hat{Q}^* = D_r^{1/2}\hat{B}D_c^{-1/2}$ 라는 관계에서 출발하여, 행 프로파일의 중심과의 차이인 **B**라는 데이터(행렬)를 환경변수 X_s의 공간으로 회귀 추정한 후, 열 표준화한 $\hat{B}D_c^{-1/2}$ 행렬에, 행 가중치 $D_r^{1/2}$가 부여된 가중데이터 의미를 지니고 있다. 또한, $Q = D_r^{-1/2}(P - \underline{r}\,\underline{c}')D_c^{-1/2}$라는 관계에서 보면, $\hat{Q}$은 각 셀에 대해 환경인자 X_s가 고려되어 추정된 후의 독립성 검정을 위한 각 셀의 의미를 아울러 지닌다. 최종적으로 Legendre's CCA는 Benzecri's CA와 동일하게 $\hat{B}D_c^{-1/2}$를 $D_c^{-1/2}v$ 축에 사영하는 것이다. 이를 위해 $\hat{Q} = D_r^{1/2}\hat{B}D_c^{-1/2}$, $\tilde{v} = D_c^{-1/2}v$로 놓으면,

$$\hat{Q}'\hat{Q} = \left(D_r^{1/2}\hat{B}D_c^{-1/2}\right)'\left(D_r^{1/2}\hat{B}D_c^{-1/2}\right) = \tilde{V}D_\lambda\tilde{V}'$$

라는 고유치분해로 기저벡터 v와 사영값 $\hat{B}D_c^{-1}v$를 유도할 수 있다. Benzencri's 대응분석에서 $B_z(= Q)$를 특이치 분해하듯 동일한 방법으로 $\hat{Q}$을 특이치분해(SVD)하기로 하자.

$$\hat{Q} = D_r^{1/2}\hat{B}D_c^{-1/2} = \tilde{U}D_{\sqrt{\lambda}}\tilde{V}' \tag{2.8}$$

SVD에 의해, 행(site) 사영은 $\hat{B}D_c^{-1/2}$를, 열(species) 사영은 열좌표를 각각 $D_c^{-1/2}v$에 사영한 것으로 사영 결과는 다음과 같다.

- Species Score : $D_c^{-1}v = \hat{Q}^*\tilde{U} = D_c^{-1/2}\tilde{V}$, $\quad D_c^{-1/2}\tilde{V}D_{\sqrt{\lambda}}$(주좌표 개념)
- Site Score : $\hat{B}D_c^{-1}v = \hat{Q}^*\tilde{V} = D_r^{-1/2}\tilde{U}D_{\sqrt{\lambda}}$, $\quad D_r^{-1/2}\tilde{U}$(표준좌표 개념)
- Species-environment correlation : Cor $(D_r^{1/2}X_s, \tilde{U}) = X_s'D_r^{1/2}\tilde{U}$

여기서 주좌표(principal coordination) 개념에 의해 열 사영(species score)에 $D_{\sqrt{\lambda}}$가 붙는다.

이상의 과정은 $Q = D_r^{-1/2}(P - \underline{r}\,\underline{c}')D_c^{-1/2}$에 대한 대응분석과 사실상 동일함을 보여주고 있다. 유일한 차이점은 환경인자를 사용하여 추정한 $\hat{Q}$에 대해 Benzecri's 대응분석을 적용한다는 점이다.

Table 2: Camparison of TCCA and LCCA

Method	Ter Braak's CCA	Legendre's CCA
Fundamental matrix	$(X_s'D_rX_s)^{-1/2}(X_s'P)D_c^{-1/2}$	$D_r^{1/2}X_s(X_s'D_rX_s)^{-1}X_s'(P-\underline{r}\,\underline{c}')D_c^{-1/2}$
Size	$q \times p$	$n \times p$
Rank	$\sum \text{diag}\,[T_b^G T_b]$	$\sum \text{diag}\,[\hat{Q}^G\hat{Q}]$
SVD	$U_t D_{\sqrt{\lambda}} V_t'$	$\tilde{U}D_{\sqrt{\lambda}}\tilde{V}'$
$\hat{\beta}$	$(X_s'D_rX_s)^{-1/2}U_t$	$(X_s'D_rX_s)^{-1}X_s'(P-\underline{r}\,\underline{c}')D_c^{-1/2}$
Site Score	$Z = X_s(X_s'D_rX_s)^{-1/2}U_t$	$D_r^{-1/2}\tilde{U}$
Species Score	$D_c^{-1/2}V_tD_{\sqrt{\lambda}}$	$D_c^{-1/2}\tilde{V}D_{\sqrt{\lambda}}$
Correlation	Cor $(D_r^{1/2}X_s, D_r^{1/2}Z) = X_s'D_rZ$	Cor $(D_r^{1/2}X_s, \tilde{U}) = X_s'D_r^{1/2}\tilde{U}$

[참고] T_b^G와 $\hat{Q}^G$는 각각 T_b와 $\hat{Q}$의 일반화역행렬이다.

3. 두 모형의 동일성

가우시안 반응모형을 사용한 TCCA와 가중주성분분석을 활용하는 LCCA의 동일성을 위해 기초행렬에 대한 고유벡터, 선형결합점수, 그리고 상관계수를 비교하기로 한다(참고: Table 2).

3.1. 기초행렬의 동일성

TCCA는 $T_b = (X_s'D_rX_s)^{-1/2}(X_s'P)D_c^{-1/2}$를 특이치 분해하며, LCCA는 $\hat{Q} = D_r^{1/2}X_s(X_s'D_rX_s)^{-1}X_s'D_r^{1/2}Q$ 를 특이치 분해한다. 이에 따라, 두 방법의 동일성을 따지기 위해서, 각각의 $T_b'T_b$와 $\hat{Q}'\hat{Q}$의 고유치 분해(행렬의 동일성)가 동일한가를 확인하기로 한다. 여기서 $Q = D_r^{-1/2}(P-\underline{r}\,\underline{c}')D_c^{-1/2}$라는 사실을 사용하면, $\hat{Q}'\hat{Q}$ 는 다음과 같다.

$$
\begin{aligned}
\hat{Q}'\hat{Q} &= D_c^{-1/2}(P-\underline{r}\,\underline{c}')'X_s(X_sD_rX_s)^{-1}X_s'D_r^{1/2}D_r^{1/2}X_s(X_sD_rX_s)^{-1}X_s'D_r^{1/2}D_r^{-1/2}(P-\underline{r}\,\underline{c}')D_c^{-1/2}\\
&= D_c^{-1/2}(P-\underline{r}\,\underline{c}')'X_s(X_s'D_rX_s)^{-1}(X_s'D_rX_s)(X_sD_rX_s)^{-1}X_s'(P-\underline{r}\,\underline{c}')D_c^{-1/2}\\
&= D_c^{-1/2}(P-\underline{r}\,\underline{c}')'X_s(X_s'D_rX_s)^{-1}X_s'(P-\underline{r}\,\underline{c}')D_c^{-1/2}\\
&= D_c^{-1/2}P'X_s(X_s'D_rX_s)^{-1}X_s'PD_c^{-1/2} - D_c^{-1/2}P'X_s(X_s'D_rX_s)^{-1}X_s'\underline{r}\,\underline{c}'D_c^{-1/2}\\
&\quad - D_c^{-1/2}\underline{c}\,\underline{r}'X_s(X_s'D_rX_s)^{-1}X_s'PD_c^{-1/2} + D_c^{-1/2}\underline{c}\,\underline{r}'X_s(X_s'D_rX_s)^{-1}X_s'\underline{r}\,\underline{c}'D_c^{-1/2} \qquad (3.1)\\
&= D_c^{-1/2}P'X_s(X_s'D_rX_s)^{-1}X_s'PD_c^{-1/2} \quad \text{by } X_s'\underline{r} = \underline{r}'X_s = 0 \qquad (3.2)
\end{aligned}
$$

이다. 위 식 (3.1)에서 $X_s'\underline{r} = \underline{r}'X_s = 0$ 이기 때문에 첫 번째 항을 제외한 나머지 3항이 모두 0이 된다. 이제, TCCA의 기초행렬 T_b에 대해

$$
\begin{aligned}
T_b'T_b &= D_c^{-1/2}P'X_s(X_s'D_rX_s)^{-1/2}(X_s'D_rX_s)^{-1/2}X_s'PD_c^{-1/2}\\
&= D_c^{-1/2}P'X_s(X_s'D_rX_s)^{-1}X_s'PD_c^{-1/2} \qquad (3.3)
\end{aligned}
$$

이다. 위의 식 (3.2)와 식 (3.3)을 통해 두 행렬이 서로 동일함을 확인할 수 있다. 결국 $T_b'T_b$와 $\hat{Q}'\hat{Q}$의 고유치 분해 결과인 V_t와 $\tilde{V}$는 서로 동일하며, 고유값 역시 동일함을 볼 수 있다. 이에 따라 두 방법의 종점수(species score) $D_c^{-1/2}V_t = D_c^{-1/2}\tilde{V}$ 역시 서로 동일하다 두 방법 모두 종을 주좌표로 고려하기에 종 점수에 $D_{\sqrt{\lambda}}$를 곱하여 준다. 두 행렬의 차이점은 TCCA는 P를, LCCA는 $(P-\underline{r}\,\underline{c}')$을 사용한다는 점이다. Benzecri's 대응분석이 $D_r^{-1/2}(P-\underline{r}\,\underline{c}')D_c^{-1/2}$를 사용하고, Hayashi의 수량화 3법이 $D_r^{-1/2}PD_c^{-1/2}$를 사용하는 것처럼, CCA에서도 그 관계가 놀라울 정도로 유사하다. Benzecri's 대응분석에서는 고유벡터의 마지막 열이 0인 벡터로, 수량화 3법에서는 첫 번째 열이 1인 벡터로 계산될 뿐 실질적인 계산 차이가 전혀 없다 (Kim과 Jeong, 2013). 이에 반해, Ter Braak's CCA와 Legendre's CCA는 서로 동일한 고유벡터를 제공한다.

Table 3: Example: simulated data

Site	Species abundance table			Environmental variable	
	SP1	SP2	SP3	X1	X2
1	1	7	4	1	1.2
2	3	1	1	2	1.4
3	8	5	4	3	1.3
4	1	1	2	4	1.6

3.2. 선형결합점수(linear combination score)의 동일성

LCCA의 $\tilde{U} = \hat{Q}\tilde{V}D_{\sqrt{1/\lambda}}$, $\tilde{V} = D_c^{-1/2}V$ 등에 의해, LCCA의 장소점수는 다음과 같이 전개된다.

$$\begin{aligned}
D_r^{-1/2}\tilde{U} &= D_r^{-1/2}D_r^{1/2}X_s(X_s'D_rX_s)^{-1}X_s'(P-\underline{r}\,\underline{c}')D_c^{-1/2}\tilde{V}D_{\sqrt{1/\lambda}} \quad \text{by (2.7)}\\
&= X_s(X_s'D_rX_s)^{-1}X_s'(P-\underline{r}\,\underline{c}')D_c^{-1/2}\tilde{V}D_{\sqrt{1/\lambda}}\\
&= X_s(X_s'D_rX_s)^{-1/2}(X_s'D_rX_s)^{-1/2}X_s'(P-\underline{r}\,\underline{c}')D_c^{-1/2}\tilde{V}D_{\sqrt{1/\lambda}}\\
&= X_s(X_s'D_rX_s)^{-1/2}T_b\tilde{V}D_{\sqrt{1/\lambda}}, \quad \text{where} \quad T_b=(X_s'D_rX_s)^{-1/2}X_s'(P-\underline{r}\,\underline{c}')D_c^{-1/2} \quad \text{by (3.2)}\\
&= X_s(X_s'D_rX_s)^{-1/2}U_t, \quad \text{where} \quad U_t=T_bV_tD_{\sqrt{1/\lambda}}, \tilde{V}=V_t.
\end{aligned}$$

위의 식에서 $\tilde{V} = V_t$는 동일한 계수(rank)하에서 성립한다. 결과적으로 두 모형의 장소점수는 서로 동일하며, LCCA의 장소점수가 TCCA의 선형결합(linear combination) 점수 Z로 표현됨을 확인할 수 있다.

3.3. 상관계수의 동일성

TCCA에서 환경점수와 가중 선형결합 점수와의 상관계수는 다음과 같이 표현된다.

$$\begin{aligned}
\text{Cor}\left(D_r^{1/2}X_s,\ D_r^{1/2}Z\right) &= X_s'D_rZ \quad \text{where} \quad Z=D_r^{-1/2}\tilde{U}\\
&= X_s'D_rD_r^{-1/2}\tilde{U} = X_s'D_r^{1/2}\tilde{U}\\
&= \text{Cor}\left(D_r^{1/2}X_s,\ \tilde{U}\right)
\end{aligned}$$

위 식에서 상관계수가 두 행렬의 곱으로 바로 표현될 수 있는 이유는 X_s와 Z가 서로 표준화되어 있기 때문이다. 결론적으로 TCCA와 LCCA의 상관계수 역시 서로 일치한다.

4. 수치 예 및 특징 비교

4.1. 모의자료에 대한 수치 예

Table 3은 4개의 장소에서 3 종류의 생물에 대한 출현정도와 2개 환경변수의 측정값을 보여주는 모의자료이다. 다음의 행렬 계산에 의해 기초행렬 T_b, $\hat{Q}$의 ${T_b}'T_b$와 $\hat{Q}'\hat{Q}$가 서로 동일함을 확인할 수 있다. 한편, 이들의 계수(rank)는 모두 환경변수 개수인 2로 나타났다.

$$T_b=\begin{pmatrix} 0.194 & -0.166 & -0.024 \\ 0.051 & -0.122 & 0.082 \end{pmatrix}, \quad {T_b}'T_b=\begin{pmatrix} 0.040 & -0.038 & 0.000 \\ -0.038 & 0.042 & -0.006 \\ 0.000 & -0.006 & 0.007 \end{pmatrix}$$

Table 4: Site scores and species scores by TCCA and LCCA

Site scores					Species scores				
Sites	TCCA		LCCA		Species	TCCA		LCCA	
1	1.297	0.067	−1.297	−0.067	SP1	−0.336	−0.071	0.336	0.071
2	0.259	1.491	−0.259	−1.491	SP2	0.334	−0.056	−0.334	0.056
3	−0.594	−0.907	0.594	0.907	SP3	−0.028	0.156	0.028	−0.156
4	−1.693	1.791	1.693	−1.791					

$$\hat{Q}=\begin{pmatrix} -0.145 & 0.147 & -0.008 \\ -0.041 & 0.001 & 0.044 \\ 0.103 & -0.060 & -0.045 \\ 0.084 & -0.131 & 0.057 \end{pmatrix}, \quad \hat{Q}'\hat{Q}=\begin{pmatrix} 0.040 & -0.038 & 0.000 \\ -0.038 & 0.042 & -0.006 \\ 0.000 & -0.006 & 0.007 \end{pmatrix}$$

T_b와 $\hat{Q}$의 특이치 분해는 다음과 같다.

$$T_b=\begin{pmatrix} 0.194 & -0.166 & -0.024 \\ 0.051 & -0.122 & 0.082 \end{pmatrix}=\begin{pmatrix} -0.892 & -0.451 \\ -0.451 & 0.892 \end{pmatrix}\begin{pmatrix} 0.282 & 0 \\ 0 & 0.10 \end{pmatrix}\begin{pmatrix} -0.695 & 0.717 & -0.054 \\ -0.419 & -0.342 & 0.841 \end{pmatrix}$$

$$\hat{Q}=\begin{pmatrix} -0.145 & 0.147 & -0.008 \\ -0.041 & 0.001 & 0.044 \\ 0.103 & -0.060 & -0.045 \\ 0.084 & -0.131 & 0.057 \end{pmatrix}$$

$$=\begin{pmatrix} -0.729 & -0.038 & -0.026 \\ -0.094 & -0.541 & -0.820 \\ 0.397 & 0.607 & -0.532 \\ 0.549 & -0.581 & 0.210 \end{pmatrix}\begin{pmatrix} 0.282 & 0 & 0 \\ 0 & 0.10 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 0.695 & -0.717 & 0.054 \\ 0.419 & 0.342 & -0.841 \\ 0.585 & 0.607 & 0.538 \end{pmatrix}$$

특이치 분해 결과, $\sqrt{\lambda}_1=0.282$, $\sqrt{\lambda}_2=0.01$로 고유값이 서로 같으며, V_t와 $\tilde{V}$ 역시 계수 2에서 서로 동일함을 볼 수 있다. 최종적으로 Table 2의 식에 의한 위치점수, 종 점수는 Table 4와 같으며, TCCA의 상관계수 Cor_t와 LCCA의 상관계수 Cor_L 은

$$\mathrm{Cor}_t=\begin{pmatrix} -0.998 & -0.048 \\ -0.775 & 0.631 \end{pmatrix} \quad \mathrm{Cor}_L=\begin{pmatrix} 0.998 & 0.048 \\ 0.775 & -0.631 \end{pmatrix}$$

이다. 여기서, 두 결과의 차이는 부호에서 발생하는데, 이는 V_t와 $\tilde{V}$의 부호가 서로 반대로 계산되었기 때문이다. 결국 CCA의 1축의 부호가 서로 반대로 놓이는 결과가 나왔을 뿐, 해석은 두 분석 모두 동일하다. 이로서 모의자료를 통해 두 CCA 분석이 서로 동일함을 확인할 수 있다.

4.2. 두 방법의 특징

앞의 수치예에서 가우시안 반응곡선에 기반한 TCCA와 $\hat{Q}$에 대한 가중주성분분석인 LCCA가 완전히 동일한 결과를 유도함을 볼 수 있었다. 즉, 두 분석은 서로 동일하다. 최초의 Ter Braak (1986)의 정준대응분석은 가우시안 반응곡선 모형에서 계산된 우도를 최대로 하는 종의 발현점을 찾는 연구에서 출발한다. 그러므로, 종-발현 관계를 환경측정값의 서열에 따라 가우시안 반응곡선을 그릴 수 있다면, 종과 장소간의 관계를 이해하는데 Ter Braak's CCA 방법이 도움이 된다. 반면, 환경변수 차원에서 종의 발현과 장소와의 대응관계를 탐색하는데는 Legendre's CCA가 도움이 된다.

한편, Legendre's CCA 방법은 확장성을 지니고 있다. 이는 종-발현 정보에 더하여 환경정보가 주어진 상황에서, 환경정보로 설명될 수 있는 종-발현 정보에 대해 다양한 다변량 모형을 적용해 볼 수 있는 단초를 제공하기 때문이다. 실제로 중복분석(redundancy analysis)은 추정된 종-발현에 대한 단순 주성분분석에 해당한다. 또한, 종-발현을 다변량회귀분석으로 설명하기보다 복잡한 선형모형으로 종-발현 정도를 추정한 이후에 다양한 다변량분석을 적용하는 연구 역시 시도되고 있다 (Makarenkov와 Legendre, 2002). 그리고 Legendre's CCA는 대응분석에 기초하기 때문에 대응분석의 주좌표와 표준좌표 개념을 적용할 수 있다는 잇점이 있다. 다음 Table 5는 종(species)와 위치(site) 중심 스케일에 따른 대응관계를 보여준다. 결론적으로 TCCA와 LCCA는 동일한 결과를 유도하지만, 생태학적 관점이 강조된 분석은 TCCA, 차원축소라는 통계학적 관점이 강조된 분석은 LCCA라 할 수 있다.

Table 5: Scalings in canonical correspondence analysis

Scaling type	Species scores	Site scores
Species scaling	$D_c^{-1/2}\tilde{V}D_{\sqrt{\lambda}}$	$D_r^{-1/2}\tilde{U}$
Site scaling	$D_c^{-1/2}\tilde{V}$	$D_r^{-1/2}\tilde{U}D_{\sqrt{\lambda}}$
Symmetric scaling	$D_c^{-1/2}\tilde{V}D_{\lambda^{(1/4)}}$	$D_r^{-1/2}\tilde{U}D_{\lambda^{(1/4)}}$

5. 결론

정준대응분석은 Ter Braak (1986)이 소개한 이후, 종의 발현을 환경과 연관하여 연구하는 생태학 분야에서 가장 널리 활용되는 분석 중 하나로 쓰이고 있다. 특히, 정준대응분석으로 종과 장소 및 환경변수를 하나의 평면에 행렬도(biplot)로 나타낼 수 있어, 변수들 간의 관계성을 파악하는데 많은 도움을 받을 수 있다. 또한, 정준대응분석에서 파생된 다양한 유형의 모형들이 존재한다 (Makarenkov와 Legendre, 2002). 그런데 Ter Braak (1986)의 정준대응분석은 그 이론 전개과정 중에 여러 가정이 도입되며, 일반화선형모형, 다변량회귀모형, 정준상관분석, 방향도(gradient), 행렬도 등 다양한 통계학 사전지식이 요구된다. 반면, Legendr와 Legendre (2012)의 정준대응분석은 대응분석과 상당히 유사하다는 잇점이 있다.

본 연구는 Legendre's CCA와 Ter Braak's CCA가 서로 다른 가정에서 출발하였지만, 결국 동일한 모형임을 밝히는 연구이다. 한편으론, 가우시안 반응곡선으로부터 정준대응분석을 자세히 소개한 Jeong (2020)의 연구에 대한 확장이라 할 수 있다. 그리고 계량형 다차원척도법(metric MDS)이 인자분석(factor analysis)과 동일함을 보이는 연구 방법과도 일맥 상통한다 (Jeong, 2018). 이러한 과정을 통해 다양한 다변량 수량화 방법론 들이 서로서로 연관되어 있음을 알 수 있고, 이를 가능케하는 특이치분해의 묘미를 발견할 수 있다.

정준대응분석은 생태학 이외에서는 잘 알려져 있지 않지만, 사회과학 데이터에 바로 적용할 수 있으며, 정준대응분석과 유사한 연구방법론을 인자분석이나 주성분분석, 다차원척도법 등에 적용하여 새로운 정준형 모형을 도출할 수 있으리라 예상한다. 향후 많은 연구를 통해 다양한 다변량 정준형모형이 개발되길 기대한다.

References

Aart PJM and Smeenk EN (1975). Correlations between distributions of hunting spiders and environmental characteristics in a dune area, *Netherlands Journal of Zoology*, **25**, 1–45.

Benzecri JP, collaborateurs (1973). *Analyse des Données, Tôme 1: La Classification, Tome 2: 'Analyse des Correspondances*, Paris:Dunod, France.

Hegde LM and Naik DN (1999). Canonical correspondence analysis in SAS software. In *Proceeding of the*

twenty-fourth annual SAS Users Group International (SUGI) conference, Paper 278, 1607–1613.

Hegde LM (2012). *Implementing and Interpreting Canonical Correspondence Analysis in SAS*, Frostburg State University Technical Report, Frostburg.

Hegde LM and Naik DN (2008). Canonical correspondence analysis : Some new interpretations and computations using SAS, *Journal of Statistics and Applications*, **3**, 277–302.

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

Huh M (1998). *Quantification Methods 1, 2, 3, and 4*, Free Academy Press, Seoul.

Huh M (2011). *Exploratory Multivariate Data Analysis*, Free Academy Press, Seoul.

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

Jeong HC (2018a). A Study of metric multidimensional scaling for ecology data, *Journal of the Korean Data Analysis Society*, **20**, 1759–1768. (in Korean).

Jeong HC (2018b). Gaussian response curves fitting for ecology data, *Journal of the Korean Data Analysis Society*, **20**, 2307–2318. (in Korean).

Jeong HC (2020). A study of canonical correspondence analysis based on Gaussian response model, *Journal of the Korean Data Analysis Society*, **22**, 1849–1861 (in Korean).

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

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

Ko H, Jhun M, and Jeong HC (2015b). Applications of bootstrap methods for canonical correspondence analysis, *The Korean Journal of Applied Statistics*, **28**, 485–494 (in Korean).

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

Makarenkov V and Legendre P (2002). Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression, *Ecology*, **83**, 1146–1161.

R Package ANACOR (2009). *Simple and Canonical Correspondence Analysis*, CRAN.

R Package VEGAN (2011). *Multivariate Analysis of Ecological Communities in R*, CRAN.

R Package ADE4 (2012). *Analysis of Ecological Data*, CRAN.

Ter Braak CJF (1986). Cannonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis, *Ecology*, **67**, 1176–1179.

Ter Braak CJF (1988). *CANOCO: A fortran program for canonical community ordination*, USA.

Received August 11, 2021; Revised September 24, 2021; Accepted September 26, 2021

가중주성분분석을 활용한 정준대응분석과 가우시안 반응 모형에 의한 정준대응분석의 동일성 연구

정형철 [1,a]

[a]수원대학교 데이터과학부

요 약

본 연구에서는 가중주성분분석으로부터 정준대응분석을 유도하는 Legendre와 Legendre (2012)의 알고리즘을 고찰하였다. 그리고, 가중주성분분석에 기반한 Legendre와 Legendre (2012)의 정준대응분석이 가우시안 반응모형에 기초한 Ter Braak (1986)의 정준대응분석과 동일함을 다루었다. 생태학에서 종의 발현 정도를 잘 설명할 수 있는 가우시안 반응곡선에서 도출된 Ter Braak (1986)의 정준대응분석은 종 패킹 모형(species packing model)이라는 기본 가정을 사용한 후 일반화선형모형과 정준상관분석을 결합시키는 방법으로 도출된다. 그런데 Legendre와 Legendre (2012)의 알고리즘은 이러한 가정없이 Benzecri의 대응분석과 상당히 유사한 방법으로 계산되는 특징을 지닌다. 그러므로 가중주성분석에 기초한 정준대응분석을 사용하면, 결과물 활용에 약간의 유연성을 지닐 수 있게 된다. 결론적으로 본 연구에서는 서로 다른 모형에서 출발한 두 방법이 장소점수(site score), 종 점수(species score) 그리고 환경변수와의 상관관계가 서로 동일함을 보인다.

주요용어: 가중주성분분석, 가우시안 반응곡선, 정준대응분석, 특이치분해, 선형결합점수

이 논문은 2020학년도의 수원대학교 학술진흥연구비 지원에 의한 논문임.

[1](18323) 경기도 화성시 봉담읍 와우안길 17, 수원대학교 데이터과학부. E-mail: jeong.hc@suwon.ac.kr