Paper 278

Canonical Correspondence Analysis in SAS Software

Laxman Hegde
Department of Mathematics
Frostburg State University
Frostburg, MD 21532

Dayanand Naik
Department of Math and Statistics
Old Dominion University
Norfolk, VA 23529

Ecologists analyze species-environment relations from data on biological communities and their environment. Generally, the data of occurrence or abundance of each species of a taxonomic group is collected at several sites. Also data on a set of environment variables that are important in explaining the variation in species are collected. A "site" is a basic sampling unit separated in space or time from other sites, e.g. a quadrant, a woodlot or a light trap. Canonical Correspondence Analysis (CCA) is a multivariate technique to relate composition of a species when species have bell-shaped response curve with respect to environmental gradients. Note that statisticians interpret CCA as canonical correlation analysis in standard multivariate statistical analysis. Although canonical correspondence analysis and canonical correlation analysis are closely related, there are some subtle differences between the two techniques. Ter Braak (1986) has shown a complete derivation and applications of CCA techniques. Also, Ter Braak (1988) has developed a computer program CANOCO to perform CCA and several other multivariate statistical techniques to analyze species-environmental relations. CANOCO is perhaps a widely used program by ecologists and biologists, but not so well known for the researchers in other fields. In this paper, we **review mathematical contents** of CCA as discussed in Ter Braak (1986) and develop a SAS program to perform CCA and biplots to present a species environmental relationship.

Data

Let y_{ik} represent the abundance of the k^{th} species at the i^{th} site where i = 1,2,..., n and k = 1,2, ..., m. Let z_{ij} represent the value of the j^{th} environmental variable at the i^{th} site, where j = 1,2, ..., q. Also x_i represents the unknown score (ordination axis for environment) of site i.

Assumptions

Whittaker (1956, 1967) showed that species generally follow unimodal relationships with environmental variables. Gauch and Whittaker (1972) used the Gaussian curve (Fig. 1) to model the mean response level (μ_{ik}) of a species with respect to an environmental variable. That is, the species abundance y_{ik} (k^{th} species at site i) is distributed with mean

$$\mu_{ik} = c_k e^{-\frac{(x_i - u_k)^2}{2t_k^2}} \qquad (1)$$

where

x_i is an environmental gradient variable at site i or site score,
u_k is the value of x_i that gives the maximum abundance (optimum value),

c_k is the maximum abundance (c),
t_k is tolerance (like standard deviation), a measure of ecological amplitude.

It is further assumed that y_{ik}'s are Poisson random variables with mean μ_{ik} as in (1). Ter Braak (1986) made use of the following additional assumptions to construct CCA algorithm.

C1: the species' tolerances are equal to t.
C2: the species' maxima are equal to c.
C3: the species' optima (u_k) are homogeneously distributed over an interval A that is large compared to t.
C4: the site scores (x_i) are homogeneously distributed over a large interval B that is contained in A.

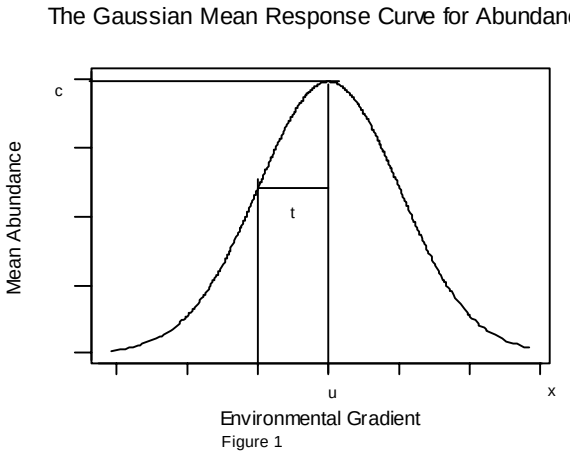

Figure 1

For easy interpretations, we standardize each of the q environmental variables as follows.

$$\sum_i w_i^* z_{ij} = 0 \text{ and } \sum_i w_i^* z_{ij}^2 = 1 \text{ where } w_i^* = \frac{y_{i+}}{y_{++}}, \quad y_{i+} = \sum_j y_{ij}, \text{ and } y_{++} = \sum_i \sum_j y_{ij}$$

(site total) (grand total)

We now model, assuming z's are standardized to have weighted mean of zero and weighted standard deviation of one, as above,

$$x_i = \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + ... + \beta_q z_{iq} = z_i' \beta \qquad (2)$$

where $z_i = \begin{pmatrix} z_{i1} \\ . \\ z_{iq} \end{pmatrix}$ and $\beta = \begin{pmatrix} \beta_1 \\ . \\ \beta_q \end{pmatrix}$, i = 1,2,...,n.

The parameters of interest are u_k, k =1,2, ..., m and β_j, j = 1,2,...,q.

Under the above assumptions, it can be shown that the maximum likelihood equations for u_k and β_j are

$$u_k = \sum_i \frac{y_{ik}}{y_{+k}} x_i - \sum_i \frac{(x_i - u_k)\mu_{ik}}{y_{+k}} \tag{3}$$

$$\sum_i z_{ij}\left[\sum_k y_{ik}(x_i - u_k)\right] = \sum_i \left[\sum_k (x_i - u_k)\mu_{ik}\right] z_{ij} \tag{4}$$

Here $y_{+k} = \sum_i y_{ik}$, $k = 1, 2, \ldots, m$, are species totals.

Under the conditions C1-C4 and equations $\sum_i (w_i^*)x_i = 0$ and $\sum_i (w_i^*)x_i^2 = 1$, we may use the approximations (Ter Braak, 1986)

$$\sum_k (x_i - u_k)\mu_{ik} \approx 0 \tag{5}$$

$$\sum_i (x_i - u_k)\mu_{ik} \approx -\lambda^* u_k y_{+k} \tag{6}$$

since μ_{ik} is symmetric about x_i and u_k; λ^* is a proportionality constant. Hence using the equation (6) in (3), we get

$$u_k = \sum_i \frac{y_{ik}}{y_{+k}} x_i + \lambda^* u_k \quad \text{or} \quad \lambda u_k = \sum_i \frac{y_{ik}}{y_{+k}} x_i \quad \text{where } \lambda = 1 - \lambda^*.$$

Next, substitute (5) in (4) to get

$$\sum_i z_{ij}\left[\sum_k y_{ik}(x_i - u_k)\right] = 0, \text{ or}$$
$$\sum_i z_{ij}\left[\sum_k y_{ik}x_i - \sum y_{ik}u_k\right] = 0, \text{ or}$$
$$\sum_i z_{ij}[x_i y_{i+} - y_{i+}x_i^*] = 0,$$

where $y_{+k} = \sum_i y_{ik}$ and $x_i^* = \sum_k \frac{y_{ik}}{y_{i+}} u_k$.

Hence,

$$\sum_i z_{ij} x_i y_{i+} = \sum z_{ij} y_{i+} x_i^*.$$

But $x_i = \sum_j z_{ij}\beta_j$. Therefore, using matrix notations,

$$\beta = (Z'RZ)^{-1}Z'Rx^*,$$

where $x^* = \begin{pmatrix} x_1^* \\ . \\ x_n^* \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 \\ . \\ \beta_q \end{pmatrix}$, $Z = \begin{pmatrix} z_{11} & . & z_{1q} \\ z_{21} & . & z_{2q} \\ . & . & . \\ z_{n1} & . & z_{nq} \end{pmatrix}$, and

$R = \text{diag}(y_{1+}, y_{2+} ..., y_{n+})$. We also have $x = Z\beta$.

Thus we have the following equations

$$\lambda u_k = \sum_i \frac{y_{ik} x_i}{y_{+k}} \tag{7}$$

$$x_i^* = \sum_k \frac{y_{ik}}{y_{i+}} u_k \tag{8}$$

$$\beta = (Z'RZ)^{-1}Z'Rx^* \tag{9}$$

$$x = Z\beta \tag{10}$$

Starting from equation (9), we substitute for x^* from (4), u_k from (7), and finally x_i from (10) and obtain

$$(S_{21}S_{11}^{-1}S_{12} - \lambda S_{22})\beta = 0 \text{ or } (S_{22}^{-1}S_{21}S_{11}^{-1}S_{12} - \lambda I)\beta = 0, \tag{11}$$

where $S_{21} = Z'Y$, $S_{12} = Y'Z$, $S_{11} = \text{diag}(y_{+1}, y_{+2}, ..., y_{+m})$, $S_{22} = Z'RZ$, and $Y = (y_{ik})$ is the m by m matrix of species abundance.

Similarly, successive substitutions in (7) leads to

$$(S_{12}S_{22}^{-1}S_{21} - \lambda S_{11})u = 0 \text{ or } \quad (S_{11}^{-1}S_{12}S_{22}^{-1}S_{21} - \lambda I)\mathbf{u} = 0. \tag{12}$$

But (11) and (12) are eigenvector equations in standard canonical correlation analysis. Note that the equations (11) and (12) can be solved using the singular value decomposition (SVD) of the m by q matrix $W = S_{11}^{-\frac{1}{2}}S_{12}S_{22}^{-\frac{1}{2}}$. That is, if the rank of W is r ($\leq$ q),

$$W = PDQ^T, \; P^TP = Q^TQ = I, \tag{13}$$

where columns of m by r matrix P are called the left singular vectors of W, the columns of r by q matrix Q are called the right singular vectors of W, and D is the r by r diagonal matrix of singular values, ordered from largest to smallest. The left singular vectors P in (13) are the eigenvectors of $WW^T = S_{11}^{-\frac{1}{2}}S_{12}S_{22}^{-1}S_{21}S_{11}^{-\frac{1}{2}}$ which solves (12) with $u = S_{11}^{-\frac{1}{2}}P$. Similarly the right singular vectors Q in (13) are the eigenvectors of W^TW =

$S_{22}^{-\frac{1}{2}} S_{21} S_{11}^{-1} S_{12} S_{22}^{-\frac{1}{2}}$ which solves (11) with $\beta = S_{22}^{-\frac{1}{2}} Q$. **Hence W is the fundamental matrix for CCA**. Once the solutions are obtained to the equations (11) and (12), ecologists are interested to study sample scores, species scores, species-environmental correlations, and correlations of environmental variables with ordination axes. These ideas are presented below.

Sample scores are computed as a n by r matrix $X = ZB$ where B is the matrix of first r eigenvector solutions ($\beta_i = S_{22}^{-\frac{1}{2}} Q,\ i = 1, 2, \ldots$r) of (11) corresponding to the first r ordered eigenvalues. Given the diagonal matrix R^* with the i^{th} diagonal element being $w_i = \frac{y_{i+}}{y_{++}}$, i=1,2,...,n and J as the n by 1 vector of 1's, the matrix X is standardized to X such that $X' R^* J = 0$ and the diagonal elements of $X' R^* X$ are one's. The columns of X in CANOCO are called as "**sample scores as linear combinations of environmental variables**". Furthermore, the q by r solution matrix of the equation $ZB = X$, say $\hat{B}$, is called the matrix of canonical coefficients corresponding to the first r eigen axes. Ecologists are also interested in knowing the correlation of an environmental variable with an ordination axis. This is computed as weighted correlation matrix of Z with X, weights being the diagonal elements of R^*. Since Z and X are standardized matrices, the q by r correlation matrix is $Z' R^* X$.

Species scores matrix U (m by r) is computed as $U = S_{11}^{-1} Y' X D^{-\alpha}$ where D is r by r diagonal matrix of eigenvalues of (11) and α = 0 or 0.5 or 1. Note that species scores represent the species optima as described in our model equation (1). These species scores are further utilized to compute n by r matrix of scores $X^* = R^{-1} Y U D^{\alpha-1}$. In Ter Braak (1988), X^* is called simply as "**species scores**". The weighted correlation of X with X^* is called as "species-environment correlation", the weights are the diagonal elements of R. If the columns of X^* are standardized to have weighted mean of zero and weighted standard deviation of 1, then species-environment correlation can be computed as r by r matrix $X' R X^*$. In the following discussions, we will review the necessary ingredients to construct biplots.

Generalized SVD

If $\underset{mxm}{\Omega}$ and $\underset{qxq}{\Phi}$ are given positive-definite symmetric matrices, then any matrix A of rank r can be expressed as

$$\underset{mxq}{A} = \underset{mxr}{N}\ \underset{rxr}{D}\ \underset{rxq}{M^T} \qquad (14)$$

where the columns of N and M are orthonormalized with respect to Ω and Φ respectively. That is $N^T \Omega N = M^T \Phi M = I$. The columns of N and M may be called generalized left and right singular vectors respectively. The elements of the diagonal matrix D may be called generalized singular values, ordered from largest to smallest. Generalized SVD of A is achieved using ordinary SVD of $\underset{mxq}{W} = \Omega^{\frac{1}{2}} A \Phi^{\frac{1}{2}} = PDQ^T$, $P^T P = Q^T Q = I$. Then letting

$N = \Omega^{-\frac{1}{2}}P$ and $M = \Phi^{-\frac{1}{2}}Q$, we will obtain (14). Generalized SVD is used in constructing the biplots by appropriately choosing A, Ω, and Φ. See Greenacre (1984, p. 344)**.**

Biplot

A biplot is a graphical presentation of a data matrix (A) by two sets of plots, overlaid on the same coordinate system, one plot representing the rows of A and the other plot representing the columns of A. Given a generalized SVD of A as in (13), the coordinates for row vectors (plot I) are provided by the rows of the first two columns of the matrix $F = ND^{\alpha}$ and the coordinates for column vectors are provided by the rows of the first two columns of the matrix $G = MD^{1-\alpha}$ (α is usually 0, $\frac{1}{2}$, and 1). In CCA, we can consider the following four different cases for a biplot presentation of species-environmenetal relationship. For more discussions on biplots, see Khattree and Naik (1998).

Case (i): $A = S_{12}S_{22}^{-1}$, $\Omega = S_{11}^{-1}$, $\Phi = S_{22}$, then it follows that
$W = S_{11}^{-\frac{1}{2}}S_{12}S_{22}^{-\frac{1}{2}}$, and $N = S_{11}^{\frac{1}{2}}P$ and $M = S_{22}^{-\frac{1}{2}}Q$.

Case (ii): $A = S_{11}^{-1}S_{12}$, $\Omega = S_{11}$, $\Phi = S_{22}^{-1}$, then it follows that
$W = S_{11}^{-\frac{1}{2}}S_{12}S_{22}^{-\frac{1}{2}}$, and $N = S_{11}^{-\frac{1}{2}}P$ and $M = S_{22}^{\frac{1}{2}}Q$.

Note: This is the case in Ter Braak (1986). The A matrix contains the weighted averages of environmental variables with respect to species abundance.

Case (iii) $A = S_{12}$, $\Omega = S_{11}^{-1}$, $\Phi = S_{22}^{-1}$, then it follows that
$W = S_{11}^{-\frac{1}{2}}S_{12}S_{22}^{-\frac{1}{2}}$, and $N = S_{11}^{\frac{1}{2}}P$ and $M = S_{22}^{\frac{1}{2}}Q$.

case (iv) $A = S_{11}^{-1}S_{12}S_{22}^{-1}$, $\Omega = S_{11}$, $\Phi = S_{22}$, then it follows that

$W = S_{11}^{-\frac{1}{2}}S_{12}S_{22}^{-\frac{1}{2}}$, and $N = S_{11}^{-\frac{1}{2}}P$ and $M = S_{22}^{-\frac{1}{2}}Q$.

Note: This is the standard canonical correlation analysis when both Y and Z are quantitative variables.

SAS Program

We have written a SAS program to calculate sample scores, species scores, and biplot coordinates for a given data set. Also our program will draw a biplot graph. Due to space restriction, we are unable to provide the program codes here. However, we will make it available (via email or Fax) to anyone interested. Please see the contact information below.

Laxman Hegde, Dept of Math, Frostburg State University, Frostburg, MD 21532
Phone: 301-687-4777; Fax: 301-687-4795; email: lhegde@frostburg.edu

Dayanand Naik, Dept of Math & Stat, Old Dominion University, Norfolk, VA 23529
Phone: 757-683-3894; Fax: 301-683-3885; email: dnaik@odu.edu

References

Ter Braak, C.J.F. (1986). Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. Ecology: 67:1167-1179.

Ter Braak, C.J.F. (1988). CANOCO: A Fortran program for canonical community ordination by [partial] [detrended] [canonical] correlation analysis, principal components analysis and redundancy analysis (version 2.1), Microcomputer Power, USA.

Whittaker, R. H., (1956). Vegetation of the Great Smoky Mountains. Ecological Monographs 26: 1-80.

Whittaker, R. H., (1967). Gradient Analysis of Vegetation. Biological Reviews of the Cambridge Philosophical Society 49: 207-264.

Gauch, H. G., & Whittaker, R. H., (1972). Coenocline Simulation. Ecology 53, 446-451.

Greenacre, M. J. (1984). Theory and Applications of Correspondence Analysis. Academic Press, London, England.

Khattree R., and Naik D. N., (1998). Applied Multivariate Statistics with SAS Software, 2nd ed, SAS institute Inc., Cary, NC, USA.