Vegetatio 69: 69–77, 1987

The analysis of vegetation-environment relationships by canonical correspondence analysis*

Cajo J. F. Ter Braak[1,2]**
[1]*TNO Institute of Applied Computer Science, Statistics Department Wageningen, P.O. Box 100, 6700 AC Wageningen, The Netherlands, and* [2]*Research Institute for Nature Management, P.O. Box 46, 3956 ZR Leersum, The Netherlands*

Keywords: Canonical correspondence analysis, Correspondence analysis, Direct gradient analysis, Ordination, Species-environment relation, Trend surface analysis, Weighted averaging

Abstract

Canonical correspondence analysis (CCA) is introduced as a multivariate extension of *weighted averaging ordination,* which is a simple method for arranging species along environmental variables. CCA constructs those linear combinations of environmental variables, along which the distributions of the species are maximally separated. The eigenvalues produced by CCA measure this separation.

As its name suggests, CCA is also a correspondence analysis technique, but one in which the ordination axes are constrained to be linear combinations of environmental variables. The ordination diagram generated by CCA visualizes not only a pattern of community variation (as in standard ordination) but also the main features of the distributions of species along the environmental variables. Applications demonstrate that CCA can be used both for detecting species-environment relations, and for investigating specific questions about the response of species to environmental variables. Questions in community ecology that have typically been studied by 'indirect' gradient analysis (i.e. ordination followed by external interpretation of the axes) can now be answered more directly by CCA.

Introduction

Direct gradient analysis relates species presence or abundance to environmental variables on the basis of species and environment data from the same set of sample plots (Gauch, 1982). The simplest methods of direct gradient analysis involve plotting each species' abundance values against values of an environmental variable, or drawing isopleths for each species in a space of two environmental variables (Whittaker, 1967). With these simple methods one can easily visualize the relation between many species and one or two environmental variables.

Plant species experience the conditions provided by many environmental variables; therefore one might wish to analyse their joint effects. Multiple regression can be used for that purpose. However, despite some successful applications, e.g., Yarranton (1970), Austin (1971) and Forsythe & Loucks (1972), ordinary multiple regression has never become popular in vegetation science. Reasons for this include: (1) Each species requires separate analysis, so regression analysis may require an unreasonable amount of effort. (2) Vegetation data are often qualitative, or when they are quantitative the data contain many zero values for the plots at which a species is absent. In neither case do the data satisfy the assumption of a normal error distribution that is implicit in ordinary multiple regression. (3) Relationships between species and environmental variables are generally non-linear. Species abundance is often a single-peaked (bell-

* Nomenclature follows Heukels-Van der Meijden (1983). Flora van Nederland, 20th ed.

** I would like to thank the authors of the example data sets for permission to use their data, Drs M. O. Hill and H. G. Gauch for permission to use the code of the program DECORANA, and Drs I. C. Prentice, L. C. A. Corsten, P. F. M. Verdonschot, P. W. Goedhart and P. F. G. Vereijken for comments on the manuscript.

shaped) function of the environmental variables. (4) Environmental variables are often highly correlated, and so it can be impossible to separate their independent effects. Generalized Linear Modelling (Austin *et al.*, 1984; Ter Braak & Looman, 1986) provides a solution for (2) and (3), but (1) and (4) remain. Whenever the number of influential environmental variables is greater than two or three, it becomes difficult to put results for several species together so as to obtain an overall graphical summary of species-environment relationships.

A simple method is therefore needed to analyze and visualize the relationships between many species and many environmental variables. Canonical correspondence analysis (CCA) is designed to fulfil this need. CCA is an eigenvector ordination technique that also produces a multivariate direct gradient analysis (Ter Braak, 1986). CCA aims to visualize (1) a pattern of community variation, as in standard ordination, and also (2) the main features of species' distributions along the environmental variables.

Ter Braak (1986) derived CCA as a heuristic approximation to the statistically more rigorous (but computationally fraught) technique of Gaussian canonical ordination, and also showed CCA's relation to correspondence analysis (CA), alias reciprocal averaging (Hill, 1973). In this paper a simple, alternative derivation of CCA is given starting from the method of weighted averaging (WA).

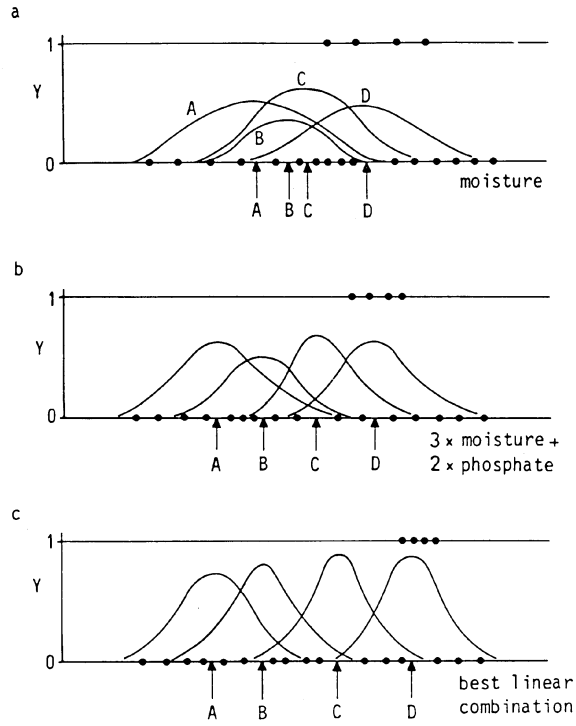

Fig. 1. Artificial example of single peaked response curves of four species (A–D) with respect to standardized environmental variables showing different degrees of separation of the species curves: (a) moisture; (b) a linear combination of moisture and phosphate, chosen apriori; (c) the best linear combination of environmental variables, chosen by CCA. Sites are shown by dots at $y = 1$ if species D is present and at $y = 0$ if species D is absent.

Theory

From weighted averaging to canonical correspondence analysis

Figure 1a shows an artificial example of single-peaked response curves for four species along an environmental variable (e.g. moisture). Species A occurs in drier conditions than species D. Fig. 1a shows presence-absence data for species D: the species is present at four of the sites.

How well does moisture explain the species' data? The fit could be formally measured by the deviance between the data and the curves, as in logistic regression (Ter Braak & Looman, 1986), but this idea will not be pursued here. Instead, a simple alternative based on the method of weighted averaging (WA) is used.

For each species a score can be calculated by taking the weighted average of the moisture values of the plots. For abundance data, this score is calculated as

$$u_k = \sum_{i=1}^{n} y_{ik} x_i / y_{+k} \qquad (1)$$

where u_k is the weighted average of the k-th (out of m) species, x_i is the (moisture) value of the i-th (out of n) site and y_{ik} is the abundance of species k at site i, and y_{+k} is the total abundance of species k. For presence-absence data the weighted average is simply the average of the moisture values of the plots in which the species is present. The weighted average