AN ALGORITHM FOR PREDICTIVE ORDINATION

László ORLÓCI*

Department of Plant Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B7

Keywords:
Linear, Non-linear, Ordination, Prediction, Vegetation

The problem

Assume that compositional variation is observed in a vegetation sample. If this variation is directed, meaning that it is not without a clearly defined trend, response to a directed environmental influence is indicated. An ordination algorithm is sought which can predict local levels of environmental influence based on observed compositional characteristics of the vegetation.

Algorithms

The predictive use of ordinations is not novel (see Orlóci 1978), but the proposed algorithms differ considerably. In one group of methods, prediction is based on simple averaging,

$$\bar{X}_j = \sum Y_{hj} X_h / Y_{.j};\ h = 1, \ldots, p \qquad (1)$$

(Whittaker 1967). $\bar{X}_j$ is a prediction of the level of environmental influence at point j of a given environmental gradient X at which the quantity of species h is Y_{hj}. The optimum of species h is at X_h. The optima $X_1, \ldots X_p$ are given a priori. $Y_{.j}$ is the sum of all Y_{hj} at X_j for all of the p species.

It is a weakness of (1) that it does not make provision for the possibility of species optima not being symmetrically dispersed about X_j on the predicted gradient, and for the unequal performance Y_{hj} of species at their optima. The predictions based on (1) can thus be very unreliable.

The algorithm, PROD, which is the subject of further discussions in the present paper, incorporates the notion of species response directly. The steps include:

1. Statement of a general hypothesis (H) which describes the exact functional form of species response.
2. Computation of ordination scores (co-ordinates for quadrats) which are completely consistent with H.
3. Interpretation of the ordination scores as predictions, considered proportional in quantity with the levels of environmental influence.

Description of PROD

In predictive ordination the computational problem is to derive t sets of co-ordinates for n points,

$$\mathbf{X} = \begin{bmatrix} X_{11} \ldots X_{1n} \\ \ldots \\ X_{t1} \ldots X_{tn} \end{bmatrix}$$

The X_{ij} are assumed to be proportional to the actual levels of environmental influence, as they are also completely consistent with the measured responses on p species,

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \ldots Y_{1n} \\ \ldots \\ Y_{p1} \ldots Y_{pn} \end{bmatrix}$$

and with a hypothesis (H) which specifies the functional form $Y_{ij} = (f(X_{ij}|m_i))$ of the trajectory of average response, conditional on a single parameter or a parameter set m_i.

Based on $f(X|m)$, the anticipated compositional distance $d(j,k|i)$, corresponding to a $\Delta(j, k|\ i) = |X_{ij} - X_{ik}|$ separation on the ith gradient is proportional to

$$d^2(j, k|i) = \int_{-\infty}^{\infty} \{f(X|m) - f(X + \Delta(j, k|i)|m)\}^2 dX \qquad (2)$$

* The project has been supported by N.S.E.R.C. of Canada funds.

The composite compositional distance is proportional to

$$d^2(j, k) = \sum d^2(j, k|i),\ i = 1, \ldots, t \quad (3)$$

H can of course be defined to apply to individual species responses and not just to an average. For the hth species, $Y_{hij} = f(X_{ij}|m_{ih})$ is the hypothesis and $d(j, k,| h, i)$ is the compositional distance. The composite compositional distance is proportional to

$$d^2(j, k) = \sum\sum d^2(j, k|\, i, h);\ h = 1, \ldots, p;\ i = 1, \ldots, t.$$

This assumes that the species responses are independent. Now if $d(j, k)$ is used for the ordination distances, the co-ordinate set $\boldsymbol{X}$ which minimizes stress,

$$\sigma^2(\boldsymbol{D};\ \boldsymbol{\delta}) = 1 - \rho^2(\boldsymbol{D};\ \boldsymbol{\delta}), \quad (4)$$

between the ordination distances ($\boldsymbol{D}$) and the observed compositional distances ($\boldsymbol{\delta}$) is the best predictive ordination possible under the given H. $\rho(\boldsymbol{D};\boldsymbol{\delta})$ measures correlation on a zero to one scale.

The function $f(X|m)$ may be of a complex shape, in which case σ may not be stationary, but strongly varying between k consecutive segments of $f(X|m)$. This can be detrimental if excessive, and it requires remedies. The solution is conceptually simple: compute a different δ_q, $q = 1, \ldots, k$, for each segment. The computational consequences may however be severe.

Examples

A linear hypothesis (LH) is rarely appropriate, but if the gradient is short, LH may be plausible. If all p trajectories have the same sense, the function takes the very simple form of

$$f(X|m) = bX + c \quad (5)$$

In any linear case, the desired predictions are obtained as the component scores in the following algorithm:

1. Compute compositional distances $\boldsymbol{\delta}$ based on

$$\delta(j,k) = [2(1 - \cos\alpha_{jk})]^{\frac{1}{2}} \quad (6)$$

where α_{jk} is the subtending angle of two quadrat vectors.

2. Extract the non-zero eigenvalues ($\boldsymbol{\lambda}$) and eigenvectors ($\boldsymbol{X}$) of Q, with elements $q_{jk} = -0.5\ (\delta^2(j, k) - \delta_j^2 - \delta_k^2 + \delta^2)$ (Orlóci 1978), and adjust the jth element in the ith eigenvector X_{ij} so that $X_{i1}^2 + \ldots + X_{in}^2 = \lambda_i$ for any i. The adjusted X_{ij} are the predictions sought.

3. Compute correlations between the t sets of predictions and the measured environmental variables to establish the identity of the most likely influential factors.

It is generally believed that a non-linear hypothesis (NH) is more appropriate than a linear one. There is however no agreement between the proponents of NH regarding the most likely functional form of the response trajectory. That this may vary depending on species and site conditions is conceivable. It is however often felt that the trajectories conform to a universal curve which is continuous (Curtis & McIntosh 1951, McIntosh 1967), and more or less bellshaped (Whittaker 1956, van Groenewoud 1965). It has been described as Gaussian for $m = (a,I)$,

$$f(X|a, 1) = exp\ -(X-a)^2/2 \quad (7)$$

Standardization is implied. Let us assume that this is indeed the case. The average response at point j on the ith gradient is given by

$$Y_{ij} = \exp\ -(X_{ij} - a_i)^2/2 \quad (8)$$

With such a trajectory for Y, the expected compositional distance is proportional to

$$d^2(j, k|i) = \frac{1}{\sqrt{\pi}} \{e - (X-a_i)^2/2 - \exp -(X + |X_{ij} - X_{ik}| - a_i)^2/2\}^2 dX$$
$$= 2(1 - S_{ijk}) \quad (9)$$

The compositional distance is a *chord* distance, since

$$S_{ijk} = e - (X_{ij} - X_{ik})^2/4 \quad (10)$$

is a similarity measure in the interval with end points zero and one.

The composite distance is given by (3). Since a_i does not appear in S_{ijk}, and since the standard deviations are assumed unity in all cases, the same H applies irrespective of species.

What has been given as (9) is what Gauch has described in 1973. He has in fact offered an algorithm to obtain a single set of predictions. Another possibility is to embed $d(j, k)$ in the Kruskal & Carmone (1972) algorithm which has the advantage of supplying not one but t independent sets of n predictions for quadrats.

Reaffirming that the objective is to obtain the predictions (which are the elements in $\boldsymbol{X}$), the steps in PROD are as follows:

1. Compute $\boldsymbol{\delta}$ with elements,

$\delta(j, k) = [2(1-\cos\alpha_{jk})]^{\frac{1}{2}}\ j < k = 2, \ldots, n$ (12)

The α_{jk} are quantities measuring subtending angles where position vectors meet in the zero origin.

2. Specify t, the number of independent environmental influences affecting compositional variation.

3. Select t sets of n arbitrary numbers to serve as first approximations for the unknown X_{ij}.

4. Compute the $d(j, k)$ values according to formulae (3), (9), (10) and examine the stress, σ, in (4).

5. If σ is stable, or less in value than a specified threshold limit, stop. If not, change the X_{ij} a little to further reduce stress and continue at step 4.

The final set of $\boldsymbol{X}$ values are the desired predictions. These have to be examined for correlations with environmental variables to establish the identity of the predicted influence.

Discussion

The Gaussian assumption underlies several formal algorithms, each capable of producing predictions ($\boldsymbol{X}$). The algorithm of Gauch (Gauch, Chase & Whittaker 1974) and that of Johnson (1973) are iterative. The latter incorporates the statistical notions of estimation. Another algorithm, devised by Ihm & van Groenewoud (1975), is based on eigenanalysis.

The advantages of the algorithm described in the present paper are several fold. First, it can be generalized to functions other than the Gaussian provided that their integral exists. Second, it relies on a distance matrix. Once the sample distances are computed, no further access is needed to the basic data. Third, it yields not one but t sets of n co-ordinates. A program is available for the linear case. One for non-linear cases is in the developmental stage.

Summary

An algorithm is described for predictive ordination. The functional form of species response is required, but it need not be Gaussian. Any integrable function is acceptable.

References

Curtis, J.T. & R.P. McIntosh. 1951. An upland forest continuum in the prairie-forest border region of Wisconsin. Ecology 32: 476–496.

Gauch, H.G. 1973. The relationship between sample similarity and ecological distance. Ecology 54: 618–622.

Gauch, H.G., G.B. Chase & R.H. Whittaker. 1974. Ordination of vegetation samples by Gaussian species distributions. Ecology 55: 1382–1390.

Groenewoud, H. van. 1965. Ordination and classification of Swiss and Canadian coniferous forests by various biometric and other methods. Ber. Geobot. Inst. ETH Stiftg. Rubel, Zurich 36: 28–102.

Ihm, P. & H. van Groenewoud. 1975. A multivariate ordering of vegetation data based on Gaussian type gradient response curves. J. Ecol. 63: 767–777.

Johnson, R. 1973. A study of some multivariate methods for the analysis of botanical data. Ph.D. thesis, Utah State Univ., Logan, Utah.

McIntosh, R.P. 1967. The continuum concept of vegetation. Bot. Rev. 33: 130–187.

Orlóci, L. 1978. Multivariate analysis in vegetation research, 2nd ed. Junk, The Hague. p. 451.

Whittaker, R.H. 1956. Vegetation of the Great Smoky Mountains. Ecol. Monogr. 26: 1–80.

Whittaker, R.H. 1967. Gradient analysis of vegetation. Biol. Rev. 42: 207–264.

Accepted 31 October 1979