



Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model

Cajo J. F. ter Braak

Biometrics, Vol. 41, No. 4 (Dec., 1985), 859-873.

Stable URL:

<http://links.jstor.org/sici?sici=0006-341X%28198512%2941%3A4%3C859%3ACAOIAA%3E2.0.CO%3B2-S>

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Biometrics is published by International Biometric Society. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/ibs.html>.

Biometrics

©1985 International Biometric Society

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2003 JSTOR

Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model

Cajo J. F. ter Braak

TNO Institute of Mathematics, Information Processing and Statistics,
P. O. Box 100, 6700 AC Wageningen, The Netherlands

SUMMARY

Correspondence analysis is commonly used by ecologists to analyze data on the incidence or abundance of species in samples. The first few axes are interpreted as latent variables and are presumed to relate to underlying environmental variables. In this paper correspondence analysis is shown to approximate the maximum likelihood solution of explicit unimodal response models in one latent variable. These models are logistic-linear for presence/absence data and loglinear for Poisson counts, with predictors that are quadratic in the latent variable. The approximation is best when the maxima and tolerances (widths) of the response curves are equal and the species' optima and the sample values of the latent variable are equally spaced. It is still fairly good for uniformly distributed optima and sample values, as shown by simulation. For the models extended to two latent variables, the approximation is often bad because of the horseshoe effect in correspondence analysis, but improves considerably in the simulations when this effect is removed as it is in detrended correspondence analysis.

1. Introduction

Correspondence analysis is a multivariate technique primarily developed for the analysis of contingency table data (Nishisato, 1980; Greenacre, 1984). However, in ecology and archaeology, correspondence analysis is commonly applied to incidence or abundance matrices (Gauch, 1982). In ecology these matrices typically record the presence/absence or abundance of species in samples, e.g., plant species in quadrats or animal species in areas. Such matrices are not transformed to m -way contingency tables "on the grounds that the data are essentially asymmetric and the absences indicate little" (Hill, 1974). Clearly a different rationale is needed for the application of correspondence analysis to incidence or abundance data. A pertinent result concerns so-called Petrie matrices (a Petrie matrix is an incidence matrix which has a block of consecutive 1's in every row and in every column, the block of the first row starting in the first column and the block of the last row ending in the last column). The result says that if a matrix can be rearranged to a Petrie matrix by a permutation of rows and columns, then this permutation is generated by the first nontrivial solution of correspondence analysis (see Hill, 1974).

Hill (1973) introduced correspondence analysis to ecology, under the name of "reciprocal averaging." He suggested the technique as a natural extension of the method of weighted averaging used in Whittaker's (1956) "direct gradient analysis." Whittaker, among others, observed that species typically show unimodal (bell-shaped) response curves with respect to environmental gradients. For example, a plant species may prefer a particular soil moisture content, and not grow at all in places where the soil is either too dry or too wet.

Key words: Correspondence analysis; Detrended correspondence analysis; Dual scaling; Ecology; Generalized linear models; Joint plot; Reciprocal averaging; Species packing model; Unfolding; Unimodal response model.

Each species is therefore largely confined to a specific interval along an environmental variable. The value most preferred by a species was termed its "indicator value" or optimum. In Whittaker's method, the indicator value of a species is estimated by taking the average of the values of the environmental variable in those samples in which the species occurs. (For quantitative data, the average is weighted by species abundance.) Conversely, with known indicator values of species, weighted averaging is used to estimate the value of an environmental variable in a sample from the species that it contained [see e.g., Kovács (1969) for an application]. Hill (1973) showed that if iterated, this process of "reciprocal averaging" converges to a solution independent of initial indicator values, namely the first nontrivial axis of correspondence analysis (see also Greenacre, 1984, §4.2). Hill's method therefore amounts to arranging samples and species along a latent variable, an activity Whittaker (1967) termed "indirect gradient analysis." After such analysis, attempts are made to identify the latent variable by comparison with known variation in the environment (Gauch, 1982). The Petrie matrix provides a deterministic example of a response model wherein the response curves are (weakly) unimodal "block functions." Unimodal models also play an important role in unfolding theory (Coombs, 1964).

In this paper, correspondence analysis is regarded as an estimation method for latent variable models and is compared with maximum likelihood under parametric unimodal response models with respect to one or two latent variables. The models considered are loglinear and logistic-linear models with predictors that are quadratic in the latent variable(s). Ter Braak and Barendregt (in press) showed that these are the only models with Poisson and binomial error, respectively, for which the weighted average of indicator values can achieve unit asymptotic efficiency with respect to maximum likelihood. The comparison gives some idea about the model that is implicitly invoked when correspondence analysis is applied to incidence or abundance data. This comparison is important because the maximum likelihood approach may be computationally too demanding for the numbers of species and samples commonly encountered in ecological research. Moreover, when the maximum likelihood approach is considered worthwhile, the results suggest that good initial estimates can be derived from correspondence analysis or, for two latent variables, from detrended correspondence analysis (Hill and Gauch, 1980).

2. Correspondence Analysis

Nishisato (1980) takes the view that correspondence analysis, alias dual scaling, assigns real numbers or "scores" to rows and columns of a table so as to optimize a particular criterion. Consider a species-by-sample matrix $\mathbf{Y} = [y_{ki}]$ ($k = 1, \dots, m; i = 1, \dots, n$) of nonnegative real numbers, denoting the presence/absence ($y_{ki} = 1$ or 0) or count of individuals of each of m species in n samples. Let $\mathbf{u} = [u_k]$ ($k = 1, \dots, m$) and $\mathbf{x} = [x_i]$ ($i = 1, \dots, n$) contain the scores for species (rows) and samples (columns), respectively. In correspondence analysis these scores are chosen so that the weighted sum of squares of the sample scores is maximum with respect to the weighted sum of squares of the sample scores within species, i.e., the criterion maximized is

$$D^2 = \sum_i y_{+i}(x_i - z)^2 / \sum_k \sum_i y_{ki}(x_i - u_k)^2, \quad (2.1)$$

where $z = \sum_i y_{+i}x_i/y_{++}$ and the subscript + denotes summation over that subscript. Maximization of D^2 will give each species a score close to the scores of those samples in which it is abundant. (An alternative interpretation of this criterion is given in Section 4.3.) With the Lagrange method of multipliers and the sample scores centred so that $z = 0$, we obtain after some rearrangement the *transition formulae* of correspondence analysis (with

$\alpha = 0$):

$$\lambda^{1-\alpha} x_i = \sum_k y_{ki} u_k / y_{+i} \quad (i = 1, \dots, n), \quad (2.2)$$

$$\lambda^\alpha u_k = \sum_i y_{ki} x_i / y_{k+} \quad (k = 1, \dots, m), \quad (2.3)$$

where λ is a real number ($0 \leq \lambda \leq 1$). The extra parameter α governs the scaling of the species scores and the sample scores with respect to one another. There are three choices of α in common usage, namely $\alpha = 0, 1$, or $\frac{1}{2}$. Criterion (2.1) leads to $\alpha = 0$. With $\alpha = 0$, the species scores u_k are weighted averages of the sample scores x_i [equation (2.3)] and the sample scores are proportional to the weighted averages of the species scores [equation (2.2)]. With $\alpha = 1$, the role of species and samples is interchanged, also in the criterion being maximized. The third choice, $\alpha = \frac{1}{2}$, is a compromise in that it treats species and sample scores in a symmetric way.

The transition formulae have more than one solution. All solutions can be obtained from the singular value decomposition of $\mathbf{R}^{-1/2} \mathbf{Y} \mathbf{C}^{-1/2}$ (see Hill, 1974) with $\mathbf{R} = \text{diag}(y_{k+})$ and $\mathbf{C} = \text{diag}(y_{+i})$. When the left and right normalized singular vectors in this decomposition are denoted by \mathbf{q}_s and \mathbf{r}_s , corresponding to singular value $\rho_s = \sqrt{\lambda_s}$ ($s = 0, 1, 2, \dots$), then the solutions are $\mathbf{u}_s = \rho_s \mathbf{R}^{-1/2} \mathbf{q}_s y_{++}^{1/2}$ and $\mathbf{x}_s = \mathbf{C}^{-1/2} \mathbf{r}_s y_{++}^{1/2}$. The solutions are the "axes" of correspondence analysis and λ_s is termed the eigenvalue of the s th axis. The maximum singular value is always 1, corresponding to the trivial solution in which all sample and species scores equal 1. The first nontrivial solution ($s = 1$) is orthogonal to the trivial solution, hence satisfies the previously applied centering $z = 0$, and maximizes the criterion D^2 with $\mathbf{u} = \mathbf{u}_1$, $\mathbf{x} = \mathbf{x}_1$, and $D^2 = 1/(1 - \lambda_1)$. Moreover, the singular value decomposition implies that the species and sample scores, \mathbf{u} and \mathbf{x} , approximate the data in a weighted least squares sense by the bilinear model (see Nishisato, 1980)

$$\frac{y_{ki} - e_{ki}}{e_{ki}} \approx u_k x_i \quad (2.4)$$

with $e_{ki} = y_{k+} y_{+i} / y_{++}$, the expectation under the assumption of row/column independence in contingency tables.

3. A Unimodal Response Model

From now on the species-by-sample matrix \mathbf{Y} will be assumed to consist either of counts y_{ki} that are independent Poisson variables with expected value μ_{ki} , or of presence/absence (1/0) data that are independent Bernoulli variables with probability μ_{ki} that the k th species is present in the i th sample. The models assumed for μ_{ki} are loglinear and logistic-linear models (Nelder and Wedderburn, 1972) in which the linear predictor is a quadratic polynomial in the latent variable x . It is convenient to write these models in the form

$$\text{link}(\mu_{ki}) = a_k - \frac{1}{2}(x_i - u_k)^2 / t_k^2, \quad (3.1)$$

where link is the logarithmic function for counts and the logistic function for the 1/0 data. In (3.1) the parameters for the k th species are a_k , the *maximum* on log or logit scale; u_k , the *mode* or *optimum* (i.e., the value of x for which the maximum is attained); and t_k , the *tolerance*, a measure of ecological amplitude. The value of the latent variable in the i th sample is x_i , which is treated as a fixed incidental parameter. Figure 1 displays an example for 1/0 data. The loglinear model is precisely the "Gaussian" response curve that is put forward by ecologists as an ideal for species responses along a gradient [see Austin (1976) and Gauch (1982) for reviews].

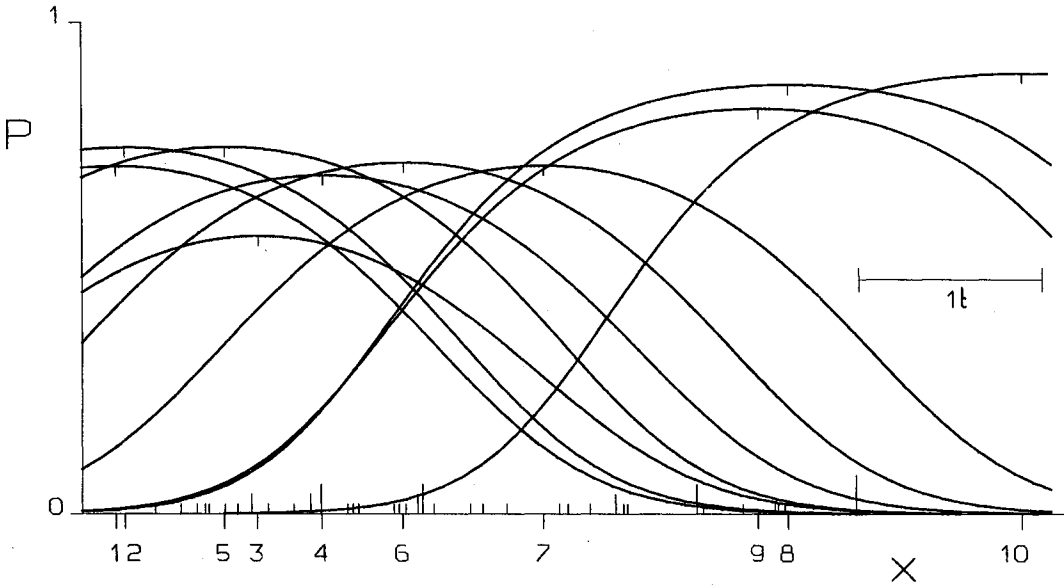


Figure 1. Unimodal response curves (3.1) for the probability (P) of occurrence along a latent variable (x), fitted by correspondence analysis to Table 2. The species optima and sample points are indicated by ticks below and above the abscissa. The length of a tick is proportional to the number of sample points. The numbers below the optima correspond to row numbers in Table 2. The horizontal bar is 1 tolerance unit.

The arbitrariness in the scale of the latent variable can be resolved, for example by centering as in correspondence analysis ($\sum_i y_{+i}x_i = 0$) and by setting the mean square of the tolerances to unity ($\sum_k t_k^2/m = 1$), so that the latent variable can be measured in (mean) tolerance units. Then, the maximum likelihood equations for the parameters $\mathbf{x} = [x_i]$ ($i = 1, \dots, n$) and $\mathbf{u} = [u_k]$ ($k = 1, \dots, m$) become, after some rearrangement,

$$x_i = \sum_k \frac{y_{ki}u_k}{t_k^2} / \sum_k \frac{y_{ki}}{t_k^2} - \left[\sum_k \frac{(x_i - u_k)\mu_{ki}}{t_k^2} / \sum_k \frac{y_{ki}}{t_k^2} \right], \tag{3.2}$$

$$u_k = \sum_i y_{ki}x_i/y_{k+} - \left[\sum_i (x_i - u_k)\mu_{ki}/y_{k+} \right]. \tag{3.3}$$

These (implicit) equations could be simplified further by using the maximum likelihood equations for the parameters $\mathbf{a} = [a_k]$ ($k = 1, \dots, m$), but for the comparison with correspondence analysis, (3.2) and (3.3) are sufficient.

4. Theoretical Comparisons

Hill's approach to correspondence analysis makes plausible that the species scores and sample scores in Section 2 play a role similar to the species optima and sample values in Section 3; that is why similar symbols are used in Sections 2 and 3. Our aim is to show that the terms between square brackets in (3.2) and (3.3) are negligible in certain cases, so that the maximum likelihood equations reduce effectively to the transitional formulae (2.2) and (2.3) of correspondence analysis. These cases are as follows: either μ_{ki} is small or μ_{ki} is symmetric around x_i and around u_k .

4.1 Equations for the Sample Scores

For the comparison of the estimation equations (2.2) and (3.2), let us first assume that x is a manifest environmental variable, and that the species' tolerances are equal ($t_k = t = 1$). With known species' optima and maxima, a missing value of the environmental variable in a sample can be estimated by using (3.1) as calibration relation. The naive estimator is the weighted average (2.2) with $\alpha = 1$. The maximum likelihood equation (3.2) would give the same result when the term between square brackets is negligible, e.g., if for all species the maximum of μ_{ki} as a function of x is close to 0 ($a_k \rightarrow -\infty$). This case may have some practical relevance, as it implies very sparse matrices, which are not uncommon in ecology.

A more interesting case arises when μ_{ki} is symmetric around x_i . This happens under the species packing model (MacArthur and Levins, 1967). This is an ecological model based on the idea that during evolution species evolve to occupy maximally separated niches with respect to a limiting resource. Christiansen and Fenchel (1977, Chap. 3) provide a lucid introduction. With x the resource, maximally separated niches mean minimal overlap between the response curves and thus, for a given number of species on a fixed-length interval and equal maxima, equal spacing between the optima (apart from edge effects). If in this situation (i) the interval is longer than, say, 10 tolerance units, (ii) the spacing between the optima on this interval is closer than ca. 1 and (iii) the sample value x_i is well within this interval, then the term between square brackets is negligible because of the symmetry in the model (3.1). Simulations showed that under the stated conditions the weighted average has, in terms of mean squared error, an efficiency of 1.00 with respect to the maximum likelihood estimator (with an uninformative prior for x_i). Moreover, Ter Braak and Barendregt (in press) showed that the asymptotic efficiency is unity when the spacing decreases to 0 on an interval of increasing length and that in the class of response curves that form a location family on x , the models considered here are the only models with this property.

The weighted average still has approximately unit efficiency when the species maxima and optima vary in a cyclic pattern along the environmental variable, i.e., when the species can be divided into sets so that within each set the species have equal maxima and equally-spaced optima with spacing less than 1 tolerance unit. However, the efficiency may drop considerably when the tolerance varies. For example, with two tolerances differing by a factor 2, the efficiency drops to ca. .6 in the logistic model with maximum probability of occurrence .5. In that case the term between square brackets still vanishes, but what remains is not a simple weighted average. If the tolerances are known a priori, then the weighted average should be applied to y_{ki}/t_k^2 , instead of to y_{ki} , in order to retain high efficiency.

More realistically, let us assume a superpopulation of response curves in which (i) the optima are independently and uniformly distributed on an interval (cf. Whittaker, Levin, and Root, 1973), (ii) the species maxima are either constant or random variables independent of the species optima, and (iii) the tolerances are equal. In this superpopulation the numerator of the term in square brackets in (3.2) vanishes in expected value, provided the sample value x_i is, again, well within the interval on which the optima are uniformly distributed. Because expectation is involved now, neglecting the term in square brackets makes weighted averaging less efficient with respect to maximum likelihood. In the logistic model with equal maxima, the asymptotic efficiencies are .96, .79, and .50 when the maximum probability of occurrence is .1, .5, and .9, respectively (Ter Braak and Barendregt, in press).

With $\alpha = 1$, the difference between the correspondence analysis equation (2.2) and the maximum likelihood equation (3.2) for latent x is the term between square brackets. The above comparisons for manifest x indicate in which situations neglecting this term does

not affect the solution too much. Note that equation (2.2) does not involve the species maxima and, further, that for equation (2.2) to be efficient for all samples, the sampled interval should be amply contained in the interval of the optima. With the choice $\alpha = 1$ the latter condition is pre-assumed.

4.2 Equations for the Species Optima

When the sample values are known a priori, estimation of the optima is a regression problem. From the symmetry between sample values and species optima in model (3.1) when the maxima and tolerances are equal, we deduce that the results of the previous section carry over to those species whose optima lie well within the sampled interval. For those species the weighted average is therefore asymptotically fully efficient with respect to the maximum likelihood estimator of the optimum, when the sample points are equally spaced with spacing less than 1 tolerance unit, and has a somewhat lower efficiency when the sample points are independently and uniformly distributed over the sampled interval (Ter Braak and Looman, in press). (That the maximum and the tolerance are to be estimated as well does not matter, because for these species the estimator for the optimum has under the stated conditions negligible correlation with the estimators for the maximum and the tolerance.) However, for species whose optima lie near the edge of, or even outside, the sampled interval, the weighted average is biased toward the center of the sampled interval, because these species' response curves are truncated. For example, the weighted average always gives a value inside the sampled interval, whereas the true optimum may lie outside this interval. This is where the eigenvalue λ of correspondence analysis comes in. With $\alpha = 1$ as in the previous section, equation (2.3) can be rewritten as

$$u_k = \sum_i y_{ki}x_i/y_{k+} - (\lambda - 1)u_k. \quad (4.1)$$

The term $(\lambda - 1)u_k$ can be considered as an overall correction term for the bias, or, alternatively, as a crude approximation to the term between square brackets in the maximum likelihood equation (3.3). The first nontrivial solution to the transition formulae has an eigenvalue λ closest to 1 and is therefore the solution where the least correction is required. This must be the solution with the longest underlying gradient, because the edge effects that cause the bias decrease with increasing length of the sampled interval. Although the correction term acts in the right direction, it overcorrects for optima well within the sampled interval and still undercorrects for optima on the edge of or outside the sampled interval. This observation explains the "compression of the first axis' ends relative to the axis middle" (Gauch, 1982) in correspondence analysis.

4.3 Scaling of the Correspondence Analysis Solution

The choice of α in the transition formulae (2.2) and (2.3) affects the scaling of the species scores with respect to the sample scores. If the sampling interval is contained well within the interval of the species optima, then α should naturally be 1 (§4.1). If the converse applies, then α should be 0. In practice, the intervals may coincide or may only partly overlap. The choice of α is then arbitrary and should be decided upon by other means (see §6.2).

The standardization of the sample scores also requires attention. Commonly the dispersion s^2 of the sample scores, $s^2 = \sum_i y_{+i}x_i^2/y_{++}$, is set equal to the eigenvalue λ , so that differences between sample scores approximate "chi-squared distances" between samples (see, e.g., Greenacre, 1984, p. 82). In the maximum likelihood approach (§3), the mean squared tolerance is set to unity. Assuming the loglinear model and the species packing

model, Hill (1979) estimated the mean squared tolerance by $\sum_k \sum_i y_{ki}(x_i - u_k)^2/y_{++}$ and standardized the correspondence analysis solution so that this estimator becomes 1. Hill's standardization gives as dispersion of the sample scores $1/(1 - \lambda)$ for $\alpha = 0$ (see §2) and $\lambda/(1 - \lambda)$ for $\alpha = 1$. Under the species packing model an alternative interpretation of criterion (2.1) is therefore that correspondence analysis maximizes the dispersion of the sample scores, subject to maintaining species response curves with unit mean squared tolerances. (By contrast, principal component analysis maximizes the variance of the sample scores subject to the condition that the sample scores are a normalized linear combination of the species' abundances.)

4.4 Conclusion

In conclusion, the transition formulae of correspondence analysis approximate the maximum likelihood equations for model (3.1). For equally-spaced optima and sample points, and equal maxima and tolerances, correspondence analysis uses a rough approximation to correct for edge effects. For uniformly distributed optima and sample points a second kind of approximation is involved, namely that the expectation is taken with respect to these uniform distributions over these parts of the maximum likelihood equations that do not depend on the data y_{ki} . The equality of the species maxima does not appear to be a crucial assumption. For unequal and unknown tolerances the approximation is worse, because the transition formulae then need to be weighted as well by the tolerances, which is not done in correspondence analysis.

5. Two Latent Variables

5.1 A Unimodal Model

The obvious extension of model (3.1) with equal tolerances to two latent variables is

$$\text{link}(\mu_{ki}) = a_k - \frac{1}{2}(x_{i1} - u_{k1})^2 - \frac{1}{2}(x_{i2} - u_{k2})^2. \quad (5.1)$$

The maximum likelihood equations for x_1 , x_2 , and u_1 , u_2 are analogous to (3.2) and (3.3) and nothing new arises in the comparison with the transition formulae. However, the edge effects due to truncation are likely to be more severe in two dimensions. First, there is more edge; second, the bias of the weighted average for, say, u_{k1} will in general depend not only on u_{k1} but also, through μ_{ki} , on u_{k2} . Approximating this bias by $(\lambda_1 - 1)u_{k1}$ is thus dubious; yet only with such approximations do the maximum likelihood equations reduce to the transition formulae of correspondence analysis.

5.2 Detrended Correspondence Analysis

Hill and Gauch (1980) developed detrended correspondence analysis as a heuristic modification of correspondence analysis, designed to correct two major "faults": (i) that the ends of the first axis are often compressed relative to the axis middle (see §4.2); (ii) that the scores of the second axis frequently show a systematic, often quadratic relation with those of the first axis. The latter fault, known as the horseshoe or arch effect, can be proven to occur for certain matrices (Hill, 1974, Proposition 8; Schriever, 1983).

Hill and Gauch (1980) adopt the species packing model to remedy the compression problem. The "species turnover rate" (assumed constant) can be estimated at a point along the gradient by the dispersion of the scores of the species present in a sample at that point. Hill and Gauch therefore try to equalize the mean within-sample dispersion of the species scores at all points along the axis by rescaling the species scores [see Hill (1979) for the details]. Thereafter the sample scores are simply derived by weighted averaging.

The horseshoe effect is considered by Hill and Gauch (1980) as "a mathematical artifact, corresponding to no real structure in the data." They eliminate the horseshoe by "detrending." Detrending intends to assure that, at any point along the first axis, the mean value of the sample scores on the subsequent axes is approximately 0. To this end the first axis is divided into a number of segments and within each segment the sample scores on axis 2 are adjusted by centering them to zero mean. The program by Hill (1979) uses running segments for this purpose. This process of detrending is built into the reciprocal averaging algorithm that generates the normal correspondence analysis solution, and replaces the usual orthogonalization procedure. Subsequent axes are derived similarly by detrending with respect to each of the existing axes.

Detrended correspondence analysis has been tested on data sets simulated under the Gaussian response model in one to four dimensions and was found to recover the structure of the data well (Hill and Gauch, 1980; Gauch, Whittaker, and Singer, 1981).

6. Numerical Comparisons

6.1 Introduction

The theoretical comparisons described so far are approximate and are supplemented in this section by numerical comparisons, using simulated data sets and one real data set. The performance of correspondence analysis is judged by correlations of the sample scores with the real values and by log-likelihood.

6.2 Methods

Data were simulated under the response models (3.1) and (5.1) in one and two dimensions, respectively, using unit tolerance and equal maxima. The optima and sample points were drawn in each simulation independently from a uniform distribution on an interval and rectangle with prechosen length and sides, respectively. Ecologists refer to such simulations as coenocline and coenoplane simulations [see Gauch (1982)]. The simulations were constrained to give at least three occurrences in each sample and at least three occurrences per species, to ensure that all parameters could be estimated.

Subroutines from Hill (1979) were used to calculate the (detrended) correspondence analysis solution for the species optima and sample scores with $\alpha = 1$ and Hill's (1979) standardization (§4.3). With these scores and $t = 1$ the species maxima were estimated by maximum likelihood, analytically in case of Poisson counts (Kooijman, 1977), and numerically in case of 1/0 data. For this solution the likelihood was calculated. In this simple approach the choice of α is arbitrary, but influences the likelihood. In a second approach this problem was solved by calculating for each species the regression of the species' responses on the sample scores. This is easy because models (3.1) and (5.1) are generalized linear models (Nelder and Wedderburn, 1972). The tolerances were kept fixed to 1 in the regressions.

The maximum likelihood solution was derived by alternating "regressions" to estimate the species parameters and "calibrations" to estimate the sample parameters, the latter being centred and, in two dimensions, rotated to principal axes in each iteration (Kooijman, 1977). Thus, regression and calibration replace the simple weighted averages in the two-way averaging algorithm to derive the correspondence analysis solution. In each regression step and each calibration step the Gauss-Newton method was used with Gallant's (1975) chopping rule for stepshortening, and a primitive method that prevented parameters from iterating to infinity. As usual, it cannot be guaranteed that the overall maximum of the likelihood is found, but the algorithm is at least hill climbing. This optimization method is akin to the EM algorithm (Dempster, Laird, and Rubin, 1977), the difference being that

with the EM algorithm it is assumed that the incidental parameters are random, whereas in this paper they are treated as fixed parameters. EM maximizes therefore a marginal likelihood (Bock and Aitkin, 1981), whereas here the joint likelihood is maximized. The (detrended) correspondence analysis solutions and also, when available, the true parameter values provided the initial parameter values.

6.3 Simulation Results

Table 1 summarizes simulations of incidence matrices (A–E) and matrices with counts (F–I), the former simulated from the logistic response curves (3.1), the latter from the loglinear response surfaces (5.1), all with unit tolerance. The maximum probability of occurrence is .7 in A, B, and C, and .5 in D and E. The maximum count is either 5 (F, G, H) or 1 (I).

Table 2 shows an example of B in which the length of the sampled interval is 5 tolerance units and Figure 1 displays its correspondence analysis solution. Although some of the species scores are out of order, the correlation of the scores of samples and of species with the true values is over .9 and the deviance is even lower than under the true parameter values. Table 1 shows that in all simulations correspondence analysis performed well for the first dimension, but in simulations F–I, badly for the second dimension. Detrended correspondence analysis is comparable to correspondence analysis in one dimension (A–E), but far superior in two dimensions (F–I).

Table 1
Results of simulations of the models (3.1) and (5.1) with unit tolerance, for 1/0 data in one dimension (A–E) and for Poisson counts in two dimensions (F–I). Shown are average values of at least four simulations (first axis 1, then axis 2, if appropriate).

Simulation	A	B	C	D	E	F	G	H	I
No. of species	30	10	30	30	30	40	40	40	40
No. of samples	20	50	50	50	50	50	50	50	50
Range of u	12	6	5	5	3	10; 5	5; 5	7; 4	7; 4
Range of x	10	5	4	4	2	8; 4	4; 4	6; 3	6; 3
Value of a	1	1	1	0	0	1.6	1.6	1.6	0
No. of par.	79	69	109	109	109	218	218	218	218
df	521	431	1391	1391	1391	1782	1782	1782	1782
Eigenvalues ($\times 100$)									
CA	90	50	38	52	18	88; 63	61; 49	77; 44	81; 57
DCA	90	50	38	52	18	88; 45	61; 39	77; 34	81; 44
Deviances									
Null model	634	654	1941	1641	1936	3448	4316	4000	1477
True par.	327	483	1556	1396	1883	836	1377	1225	856
CA	308	458	1506	1289	1778	1696	1708	1958	907
DCA	292	445	1533	1324	1789	1010	1433	1194	681
CA + REGR	264	441	1475	1280	1758	1167	1320	1374	754
DCA + REGR	279	423	1495	1309	1781	775	1255	1070	642
ML	217	417	1440	1259	1739	648	1170	994	598
Correlation with true sample scores ($\times 100$)									
CA	98	90	95	95	67	97; 57	—	98; 64	96; 53
DCA	98	90	96	91	51	98; 83	—	99; 91	96; 77
ML	99	86	94	92	67	99; 95	—	99; 93	96; 77

No. = number; u = species optima; x = sample scores; par. = parameters; df = degrees of freedom; CA = correspondence analysis; DCA = detrended correspondence analysis; (D)CA + REGR = (D)CA followed by regression on (D)CA sample scores; ML = maximum likelihood.

—: Meaningless.

Table 2

Incidence matrix simulated from unimodal response curves (3.1) under condition B in Table 1. The species (rows) and samples (columns) are arranged in increasing order of the true optima and sample values, respectively.

111111111011001010100000100010000000000000000000000
11101111100111111100110010010000000000000000000000000
1100000110010110111111110010000011000000000000000000
011100111110110010111011010101011110101100000000000000
11111100100111101110011101001010001011000000010000000
001100010111011010111101111011101110111101100000000
00010101011000011111011111001000101111101100111000
000000000000000110100001001000111111011111111111111
000000000000101000001110010011011001101110111111111
00000000000000000000000000000000100010100101110100111

In two dimensions each solution of correspondence analysis showed the horseshoe, most in F and H, least in G and I. The lower the maximum of the response curves, the better correspondence analysis (D vs C and I vs H), in accordance with the theory. The simulations also confirm the observation of Hill and Gauch (1980) that correspondence analysis works more satisfactorily with square sampling regions as compared to rectangular regions (G vs F, H). In order to determine whether the success of detrended correspondence analysis is due to the rescaling of the axes or to the detrending, some tests were done with rescaling, but without detrending. These tests showed a slight, but unimportant improvement over the results of correspondence analysis. The success of detrended correspondence analysis is therefore mainly due to the detrending.

The eigenvalues showed little variation between simulations of the same type; for example, in A and F the standard deviations were below 0.05.

The estimates of the species optima can be improved by regressing each species response on the sample scores, as can be seen from the drop in the deviance (Table 1) and the increase in correlation with the true optima (not shown). The deviance after regression on the sample scores from detrended correspondence analysis was in nearly all simulations less than the deviance under the true parameters.

The maximum likelihood solution has, by definition, the lowest deviance, but does not always give the highest correlation with the true sample scores. Of the three sets of initial values used to derive the maximum likelihood solution, the true values and the values from detrended correspondence analysis gave nearly identical solutions. Starting from the correspondence analysis solution, the maximization procedure frequently became trapped in a local maximum in simulations F-I.

For statistical tests and confidence regions it is tempting to assume that deviances are chi-squared distributed. This assumption is risky in this context because the number of parameters increases with the number of observations. Indeed, the true parameter values lie outside the usual 95% confidence region in 34% of the 29 simulations of the one-dimensional model and in 12% of the 24 simulations of the two-dimensional model.

6.4 *A Real Data Set*

The real data set, taken from Van der Aart and Smeenk-Enserink (1975), concerns the distribution of twelve wolfspiders (*Lycosidae*) in a dune area and consists of their accumulated catches in 100 samples. The maximum count in the data is 189, far higher than in the simulations, but zeroes are as equally abundant as in the simulations. Correspondence analysis was applied to these data, giving .65 and .42 for the first two eigenvalues. The

sample scores of the second axis showed a clear quadratic trend with respect to those of the first axis. Removing this trend, detrended correspondence analysis resulted in a second eigenvalue of .09. This small value indicates that the second axis is unimportant for these data, which agrees with the results of Kooijman (1977), who fitted one- and two-dimensional Gaussian response models to these data by maximum likelihood.

Table 3 shows the results of loglinear regressions of the catches of the wolfspiders on the sample scores of the first axis of detrended correspondence analysis. When a quadratic term was added to the model, the deviance decreased considerably for nine of the twelve spider species. Their fitted curves are all unimodal (see Figure 2). The rescaling of the axis in detrended correspondence analysis appears advantageous for these data, as the quadratic fit with respect to the first axis of the usual correspondence analysis resulted in a 50% higher deviance. The full maximum likelihood solution (with equal tolerances) gave a deviance of 4890, 30% lower than the deviance of the quadratic model in Table 3. Yet the sample scores as estimated by maximum likelihood showed a high correlation (.95) with those of detrended correspondence analysis.

Van der Aart and Smeenk-Enserink (1975) also characterized the vegetation and the soil around 28 of the 100 pitfall traps. They state, "The sites were selected in such a way that as many biotope types as possible were represented." Interpreting the first axis of detrended correspondence analysis as a latent variable, we can therefore attempt to relate this latent variable to the measured environmental variables. A multiple regression of the first axis' scores on the logarithms of the variables soil water content, percentage of bare sand, and percentage cover by mosses accounted for 90% of the variance. All three variables contributed to this regression, as judged by *t* tests on the regression coefficients. The first axis can therefore be interpreted as a composite gradient of soil moisture and openness of the habitat. A possible explanation for these results is that wolfspiders require an open habitat for hunting purposes but, on the other hand, require moisture to avoid desiccation. Each species balances these conflicting requirements in its own way and is therefore largely confined to a specific interval along the composite gradient of soil moisture and openness. Other factors related to soil moisture or openness cannot be excluded to be operational.

Table 3

Loglinear regressions of catches of wolfspiders (k) on the sample scores (x_i) of the first axis of detrended correspondence analysis. Given are the deviance of the null model and the decreases in deviance when the loglinear model is extended successively with a linear ($b_{k1}x$) and a quadratic term ($b_{k2}x^2$). Provided $b_{k2} < 0$, the quadratic model fits Gaussian response curves with unequal tolerances [equation (3.1)]. The spiders are arranged in order of the species score of the first axis.

<i>k</i>	Model: $\log \mu_{ki} =$	Deviance b_{k0}	Successive decrease in deviance	
			+ $b_{k1}x_i$	+ $b_{k2}x_i^2$
	Wolfspider			
1	<i>Pardosa lugubris</i>	1494	1159	11
2	<i>Zora spinimana</i>	935	245	341
3	<i>Pardosa nigriceps</i>	3109	388	1490
4	<i>Trochosa terricola</i>	3671	1033	1743
5	<i>Pardosa pullata</i>	4504	427	2570
6	<i>Arctosa lutetiana</i>	315	18	149
7	<i>Aulonia albimana</i>	958	93	488
8	<i>Alopecosa cuneata</i>	1396	57	696
9	<i>Pardosa monticola</i>	4103	130	3023
10	<i>Alopecosa accentuata</i>	856	329	202
11	<i>Alopecosa fabrilis</i>	864	693	24
12	<i>Arctosa perita</i>	340	254	3
	Total	22545	4826	10740

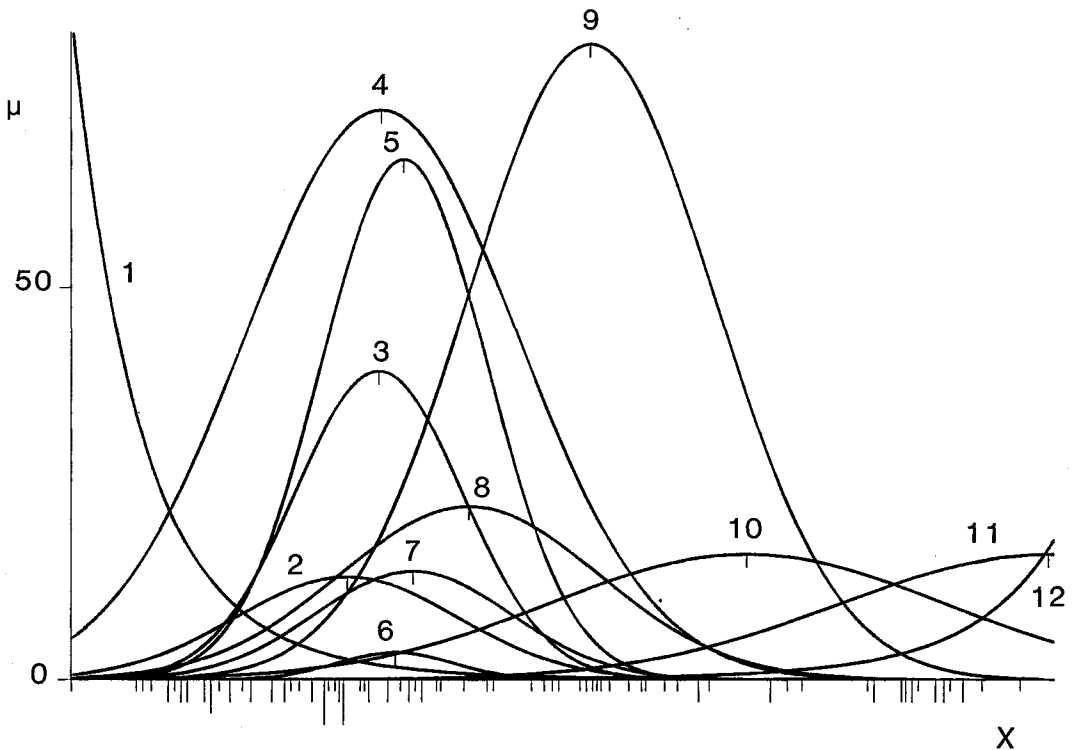


Figure 2. Unimodal response curves (3.1) for the expected number (μ) of wolfspiders along the first axis of detrended correspondence analysis (x), fitted by loglinear regression (Table 3, last column). The curves are labelled by the species identification numbers of Table 3. The sample points are indicated by ticks below the abscissa (length proportional to number). Data from Van der Aart and Smeenk-Enserink (1975).

7. Discussion

Both the unimodal model (3.1) with $t_k = t$ and the bilinear model (2.4) stand at the basis of correspondence analysis. The clue to this apparent paradox is data transformation. In linear regression, data transformation can be used to linearize *monotone* relationships. In multivariate analysis, data transformation can also be used to linearize nonmonotone relationships. Correspondence analysis is not the only example. Kooijman (1977) showed that principal component analysis recovers exactly the parameters of equal tolerance Gaussian curves and surfaces from error-free data when the data matrix is centered by rows and by columns after log transformation. Aitchison (1983) proposed this transformation to overcome the difficulty of the constant-sum constraint in principal component analysis of compositional data. He notices that "the nonlinearity of the logarithmic function opens up the possibility of coping with curvature in data sets ...," but does not refer to the Gaussian or unimodal response model. [His Figure 2(b) clearly shows the unimodal response of constituent F along the first principal component.] Ihm and Van Groenewoud (1975) used a different transformation to analyze Gaussian response curves by principal component analysis. Their method requires the same assumptions as correspondence analysis about the distribution of the optima and the sample points.

Four conditions (equal tolerances, equal or independent maxima, and equally-spaced or uniformly distributed optima and sample points) are needed to show that (detrended) correspondence analysis provides an *approximate* solution to the unimodal models (3.1) and (5.1). How realistic are these assumptions in practice and how robust is correspondence

analysis to violations of the assumptions? Some checks on the assumptions are possible, e.g., by regressing each species' responses on the derived sample scores, allowing the tolerances and maxima to vary among species, and I suggest that this should be done routinely, if only to determine the goodness-of-fit of the model for descriptive purposes. Ihm and Van Groenewoud (1975) and Kooijman (1977) reported that the optima and sample values as estimated by their methods are fairly robust against unequal tolerances, as did Hill and Gauch (1980) for detrended correspondence analysis. The four conditions are not needed in the maximum likelihood approach, taken by Gauch, Chase, and Whittaker (1974) for normal data, Kooijman (1977) for Poisson data, and Goodall and Johnson (1982) for presence/absence data. Yet, the maximum likelihood approach is applied seldom in ecological research because of its computational complexity and the lack of reliable and flexible software (Gauch, 1982). Another reason might be that correspondence analysis appears to be "nonparametric." However, this paper reveals its close connection with "Gaussian" response curves with equal tolerances.

Commonly high values in the data matrix are downweighted in correspondence analysis by, for example, a prior square root transformation. However, when the variance is proportional to the mean, transformation is not required (Wedderburn, 1974). Overdispersion then inflates the mean deviance, not necessarily implying lack of fit. When the type of dispersion or lack of fit is allowed to vary between species, all problems of common factor analysis are lurking in the way.

Principal component analysis and correspondence analysis are rival methods for dimensionality reduction for abundance data (Gauch, Whittaker, and Wentworth, 1977; Greig-Smith, 1983), both allowing "major features" of the data to be visualized in joint plots of species and sample scores. The geometrical interpretation of a principal component plot is based on the bilinear model, as stressed by Gabriel (1971), who termed the plot a *biplot*. The value of a variable, as approximated by the biplot, changes linearly across the plot. Correspondence analysis therefore gives a biplot of the transformed data values (2.4). However, in terms of the original data Y the joint plot of correspondence analysis is not a biplot, because the model for the original data is unimodal rather than bilinear. The original value of a variable, as approximated by a correspondence analysis plot, is maximum at this variable's point in the plot and decreases with distance from that point, disregarding for a moment the fact that (detrended) correspondence analysis provides only an approximate solution to the unimodal models (3.1) and (5.1). We may interpret the correspondence analysis plot more informally as Benzécri et al. (1973) do. Their centroid principle (*le principe barycentrique*) is simply the transition formulae interpreted geometrically. Multi-dimensional unfolding provides the same kind of plot (Carroll, 1972).

Although principal component analysis and correspondence analysis model and display multivariate data in different ways, the resulting plots of the sample scores are sometimes similar. This happens when all unimodal surfaces are truncated to monotone surfaces over the region actually sampled, the monotone surfaces being approximated by planes in principal component analysis. In such cases the correspondence analysis solution with $\alpha = 1$ shows some species points close to the centroid of the sample points, whereas the other species' points fall outside the region where the sample points lie.

ACKNOWLEDGEMENTS

I would like to thank Dr I. C. Prentice for valuable discussions and comments.

RÉSUMÉ

L'analyse des correspondances est couramment utilisée par les écologistes pour analyser des données de présence/absence ou d'abondance d'espèces. Les tout premiers axes sont interprétés en termes de variables sous-jacentes conditionnant la distribution des espèces. On fait l'hypothèse que ces variables

sont liées aux variables de milieu non explicitées. Dans cet article, on montre qu'en utilisant l'analyse des correspondances, on obtient une solution approchée de la solution donnée par la technique du maximum de vraisemblance dans le cas de modèles de réponse unimodale à une variable sous-jacente. Les modèles utilisés sont des modèles logistiques-linéaires en ce qui concerne les données de présence/absence et log-linéaires pour des abondances suivant des lois de Poisson, les estimateurs étant des fonctions quadratiques de la variable sous-jacente. On obtient une approximation de meilleure qualité lorsque, d'une part, les maximum et les amplitudes (tolérances des espèces aux conditions de milieu) des courbes de réponse des espèces ont mêmes valeurs et que, d'autre part, les valeurs de la variable sous-jacente correspondant aux optimum de chaque espèce et aux pointes d'échantillonnage sont régulièrement réparties. L'approximation demeure satisfaisante pour des optimum et des valeurs correspondant aux échantillons distribués uniformément, ainsi que le montre la simulation. Pour des modèles à 2 variables sous-jacentes, l'approximation est souvent mauvaise en raison de la présence d'un effet Guttman. L'approximation est de bien meilleure qualité lorsque l'on réalise des simulations après avoir retiré cet effet, ce qui se produit lorsqu'on utilise une technique d'analyse des correspondances qui efface la tendance centrale du phénomène étudié.

REFERENCES

- Aart, P. J. M. van der and Smeenk-Enserink, N. (1975). Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology* **25**, 1-45.
- Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika* **70**, 57-65.
- Austin, M. P. (1976). On nonlinear species response models in ordination. *Vegetatio* **33**, 33-41.
- Benzécri, J. P. et al. (1973). *L'Analyse des Données: II. L'Analyse des Correspondances*. Paris: Dunod.
- Bock, R. D. and Aitkin, M. A. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46**, 443-459.
- Braak, C. J. F. ter and Barendregt, L. G. (in press). Weighted averaging of species indicator values: Its efficiency in environmental calibration. *Mathematical Biosciences*.
- Braak, C. J. F. ter and Looman, C. W. N. (in press). Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio*.
- Carroll, J. D. (1972). Individual differences and multidimensional scaling. In *Multidimensional Scaling. Theory and Applications in the Behavioral Sciences. Vol. 1: Theory*, R. N. Shepard, A. K. Romney, and S. B. Nerlove (eds), 105-155. New York: Seminar Press.
- Christiansen, F. B. and Fenchel, T. M. (1977). *Theories of Populations in Biological Communities*. Berlin: Springer-Verlag.
- Coombs, C. H. (1964). *A Theory of Data*. New York: Wiley.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1-38.
- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58**, 453-467.
- Gallant, A. R. (1975). Nonlinear regression. *The American Statistician* **29**, 73-81.
- Gauch, H. G. (1982). *Multivariate Analysis in Community Ecology*. Cambridge: Cambridge University Press.
- Gauch, H. G., Chase, G. B., and Whittaker, R. H. (1974). Ordination of vegetation samples by Gaussian species distributions. *Ecology* **55**, 1382-1390.
- Gauch, H. G., Whittaker, R. H., and Singer, S. B. (1981). A comparative study of nonmetric ordinations. *Journal of Ecology* **69**, 135-152.
- Gauch, H. G., Whittaker, R. H., and Wentworth, T. R. (1977). A comparative study of reciprocal averaging and other ordination techniques. *Journal of Ecology* **65**, 157-174.
- Goodall, D. W. and Johnson, R. W. (1982). Nonlinear ordination in several dimensions. A maximum likelihood approach. *Vegetatio* **48**, 197-208.
- Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.
- Greig-Smith, P. (1983). *Quantitative Plant Ecology*, 3rd ed. Oxford: Blackwell.
- Hill, M. O. (1973). Reciprocal averaging: An eigenvector method of ordination. *Journal of Ecology* **61**, 237-249.
- Hill, M. O. (1974). Correspondence analysis: A neglected multivariate method. *Applied Statistics* **23**, 340-354.
- Hill, M. O. (1979). DECORANA—A FORTRAN Program for Detrended Correspondence Analysis and Reciprocal Averaging. Ithaca, New York: Cornell University.

- Hill, M. O. and Gauch, H. G. (1980). Detrended correspondence analysis: An improved ordination technique. *Vegetatio* **42**, 47–58.
- Ihm, P. and Groenewoud, H. van (1975). A multivariate ordering of vegetation data based on Gaussian-type gradient response curves. *Journal of Ecology* **63**, 767–777.
- Kooijman, S. A. L. M. (1977). Species abundance with optimum relations to environmental factors. *Annals of Systems Research* **6**, 123–138.
- Kovács, M. (1969). Das Corno-quercetum des Mátra-gebirges. *Vegetatio* **19**, 240–255.
- MacArthur, R. H. and Levins, R. (1967). The limiting similarity, convergence, and divergence of co-existing species. *American Naturalist* **101**, 377–385.
- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**, 370–384.
- Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and Its Applications*. Toronto: University of Toronto Press.
- Schriever, B. F. (1983). Scaling of order-dependent categorical variables with correspondence analysis. *International Statistical Review* **51**, 225–238.
- Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439–447.
- 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* **42**, 207–264.
- Whittaker, R. H., Levin, S. A., and Root, R. B. (1973). Niche, habitat and ecotope. *American Naturalist* **107**, 321–338.

Received June 1984; revised June 1985.