Ecology, 74(8), 1993, pp. 2215–2230

PUTTING THINGS IN EVEN BETTER ORDER: THE ADVANTAGES OF CANONICAL CORRESPONDENCE ANALYSIS[1]

MICHAEL W. PALMER
Department of Botany, Oklahoma State University, Stillwater, Oklahoma 74078 USA

Abstract. Canonical Correspondence Analysis (CCA) is quickly becoming the most widely used gradient analysis technique in ecology. The CCA algorithm is based upon Correspondence Analysis (CA), an indirect gradient analysis (ordination) technique. CA and a related ordination technique, Detrended Correspondence Analysis, have been criticized for a number of reasons. To test whether CCA suffers from the same defects, I simulated data sets with properties that usually cause problems for DCA. Results indicate that CCA performs quite well with skewed species distributions, with quantitative noise in species abundance data, with samples taken from unusual sampling designs, with highly intercorrelated environmental variables, and with situations where not all of the factors determining species composition are known. CCA is immune to most of the problems of DCA.

Key words: *Canonical Correspondence Analysis; Detrended Correspondence Analysis; Gradient Analysis; ordination; simulation.*

INTRODUCTION

The most common kind of data set in community ecology undoubtedly consists of the abundance or importance of taxa (usually species) indexed by sampling units (e.g., quadrats, relevés, stands, traps, etc.). Typically, these data are organized in a matrix with species as rows, sampling units as columns, and abundance (or merely presence/absence) as the entries. Since such data matrices are multidimensional, and since the human mind is limited in its capacity to visualize more than a few dimensions, ecologists are forced to find ways to extract the most important dimensions of the data set.

Fortunately, most species by sampling-unit data matrices contain much redundant information (for example, different species can respond to the same environmental gradients), and hence there are typically very few *important* dimensions (Gauch 1982*a*, *b*).

There are two basic conceptual models for analyzing species by sampling-unit matrices. One model is that in which sampling-units (hereafter referred to as *sites*, although the reader must keep in mind that sampling-units can be things other than sites, such as pitfall traps, transects, or seine samples) are arranged into (often hierarchical) groups or community types, and is known as *classification*. The other conceptual model is that in which sites and/or species can be arranged along environmental gradients, and is known as *ordination*. This paper focuses on ordination.

Ordination is increasingly used for gradient analysis, or the study of species distributions along gradients. Perhaps the most widely used ordination technique is Detrended Correspondence Analysis (DCA; Hill and Gauch 1980), which is an *indirect* gradient analysis technique. In indirect gradient analysis, environmental gradients are not studied directly but are inferred from species composition data.

Indirect gradient analysis is not circular reasoning, but rather a quite logical way to uncover factors determining community structure. It is performed regularly and intuitively by experienced field naturalists. For example, an experienced ornithologist can look at bird counts from several sites, and can (with some error) place the sites along a gradient from wet to dry, or north to south, or high elevation to low elevation even if data on these factors were absent. This is because there is pattern (and redundancy) intrinsic to the data. It is fairly simple to detect such pattern in small data sets, even for someone unfamiliar with the particular sites and species. It is, however, quite difficult to intuitively order large, complex data sets without the help of multivariate ordination techniques.

DCA has many desirable properties as an indirect gradient analysis technique. Unlike Principal Components Analysis (PCA) and Correspondence Analysis (CA), DCA does not produce the *arch* or *horseshoe effect*, a spurious second axis which is a curvilinear function of the first axis (Gauch 1982*a*, Pielou 1984, ter Braak 1985, 1987*b*, Digby and Kempton 1987). Unlike Bray–Curtis (Polar) Ordination (Bray and Curtis 1957, Beals 1984), DCA does not rely on the selection of arbitrary endpoints. Unlike Nonmetric Multidimensional Scaling (NMDS, Kruskal 1964*a*, *b*) and its variants (Sibson 1972, Minchin 1987*a*, Faith and Norris 1989, Belbin 1991), the number of dimensions

[1] Manuscript received 3 November 1992; accepted 8 February 1993.

of ordination space does not need to be specified in advance.

DCA, along with Correspondence Analysis, Canonical Correspondence Analysis (see *The Correspondence Analysis family: Canonical Correspondence Analysis*), and a few others, is a weighted averaging ordination technique. The main advantages of weighted averaging ordinations include the simultaneous ordering of sites and species (this property is shared by a few other techniques, Escoufier 1987), rapid computation (relative to NMDS), and very good performance when species have nonlinear and unimodal relationships to environmental gradients, which produces severe problems for PCA (Gauch 1982*a*, Pielou 1984, ter Braak 1985, 1986, 1987*a–d*, ter Braak and Barendregt 1986, ter Braak and Looman 1986, ter Braak and Prentice 1988).

Despite its advantages, DCA has come under increasing criticism (Beals 1984, Austin 1985, Allen 1987, Ezcurra 1987, Minchin 1987*a*, Oksanen 1987, 1988, Wartenberg et al. 1987, van Groenewoud 1992). Although some criticisms have been successfully rebutted (Peet et al. 1988), a number of problems still remain with DCA: the detrending algorithm is inelegant and arbitrary, it sometimes performs poorly with skewed species distributions, it may occasionally be unstable, it occasionally does not handle complex sampling designs very well, it may compress one end of a gradient into a "tongue" (Minchin 1987*a*, Økland 1990), and it will destroy any *true* arch that actually exists in data.

Recently, a new ordination technique, Canonical Correspondence Analysis (CCA) has come into widespread use (e.g. Stevenson et al. 1989, Whittaker 1989, Wiegleb et al. 1989, Allen and Peet 1990, Borgegård 1990, Carleton 1990, John and Dale 1990, Odland et al. 1990, Palmer 1990, Prentice and Cramer 1990, Pyšek and Lepš 1991, Retuerto and Carballeira 1991). The mathematics and models behind CCA and its variants have been most thoroughly developed by ter Braak (1985, 1986, 1987*a–d*, 1988), although others have contributed to our understanding of CCA under other names (Sabatier et al. 1989, Lebreton et al. 1991). A thorough bibliography (165 references between 1986 and 1991) of CCA and related methods has been compiled by Birks and Austin (1992).

Unlike DCA, CCA is a *direct* gradient analysis technique, and represents a special case of multivariate regression. Direct gradient analysis differs from indirect gradient analysis in that species composition is directly and immediately related to measured environmental variables. Before describing CCA in more detail, it is necessary to outline the essential features of the Correspondence Analysis family of ordination methods.

The Correspondence Analysis family

Correspondence Analysis.—CCA and DCA are both variants of Correspondence Analysis (CA). The CA algorithm can either be expressed in terms of an eigenanalysis or as a "reciprocal averaging" approach (reciprocal averaging is actually a form of eigenanalysis). The mechanics of reciprocal averaging have been described in detail elsewhere (e.g., Hatheway 1971, Hill 1974, Pielou 1984, ter Braak 1985, 1987*b*, Digby and Kempton 1987); I will give only a quick overview.

The reciprocal averaging approach is computationally simple (Fig. 1A): arbitrary numbers are assigned to each site (any nonzero numbers are acceptable; the particular numbers chosen do not influence the final outcome). These numbers are *site scores*. *Species scores* are assigned to species as the weighted average of the site scores, where the weight is the abundance of the species in each site. (This is where the data enter into the algorithm.) At this stage species scores must be re-standardized, or else scores will eventually tend towards zero. Pielou (1984) suggests standardizing from 0 to 100, ter Braak and Prentice (1988) suggest subtraction of the mean and division by the standard deviation; any linear rescaling will work. *New* site scores are assigned as the weighted average of the species scores of all species that occur in the site. Again, the weights are species abundances. The new site scores are (optionally) re-standardized. The algorithm continues reciprocally averaging (and re-standardizing) sites and species, until there is no noticeable change in species and site scores from one iteration to the next. The result is the first CA axis solution. Given a data set, an identical solution will result from any set of initial arbitrary numbers.

Computation of the second CA axis is more complicated, but is essentially the same as described above except that the linear effects of the first axis are factored out. Third and higher axes can also be readily calculated.

The reciprocal averaging algorithm has been considered by some to be "circular," "mysterious," "an art form," or "wizardry." In reality, it is merely an algorithm for eigenanalysis, one of the central techniques of matrix algebra (Pielou 1984, Digby and Kempton 1987).

The solution obtained by correspondence analysis has desirable mathematical properties. The first axis consists of the ordering of species and sites that produces the maximum possible correlation between site and species scores (Gauch 1982*a*, Pielou 1984). Second and higher axes also have maximal site–species correlation subject to the constraint that axes are orthogonal. Eigenvalues associated with each axis equal the correlation coefficient between species scores and site scores (Gauch 1982*a*, Pielou 1984). Thus an eigenvalue close to 1 will represent a high degree of correspondence between species and sites, and an eigenvalue close to zero will indicate very little correspondence. If our fundamental model of species responses to environmental gradients is unimodal (this is generally accepted; see Austin 1985, Minchin 1987*b*), then high

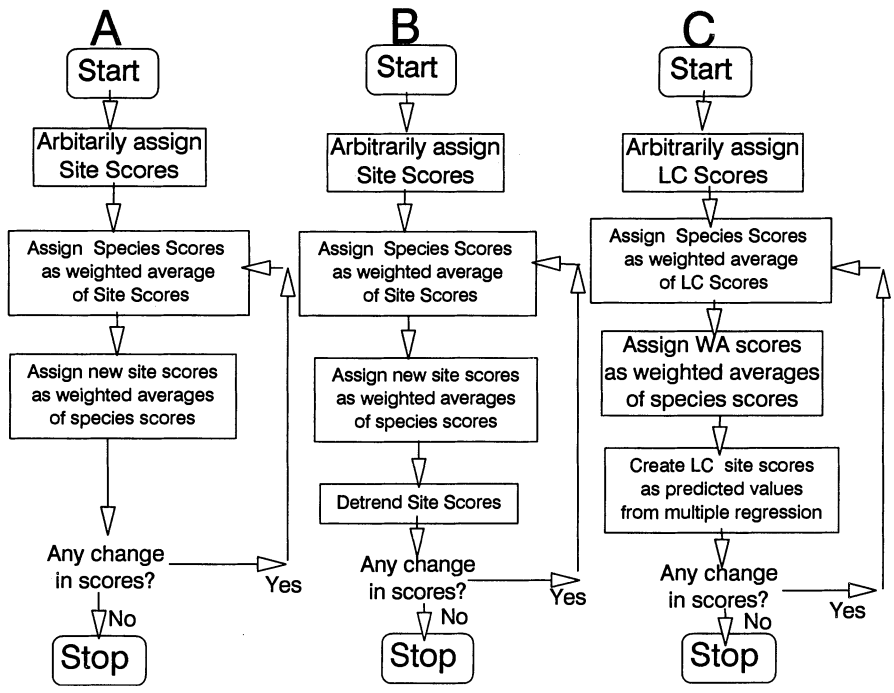

FIG. 1. Algorithms for (A) Correspondence Analysis, (B) Detrended Correspondence Analysis, and (C) Canonical Correspondence Analysis, diagrammed as flowcharts. LC scores are the linear combination site scores, and WA scores are the weighted averaging site scores.

eigenvalues are associated with long and strong environmental gradients (Gauch 1982*a*).

Detrended Correspondence Analysis.—DCA is identical to CA except that a detrending step is added (Fig. 1B). The detrending consists of removing the previously described "arch effect" by various artifices, such as cutting the first axis into segments and re-setting the average of each segment to zero (Hill and Gauch 1980), or by fitting a polynomial, usually quadratic, equation to the relationship and subtracting its effect (ter Braak 1987, Knox 1989). Site scores may also be rescaled to equalize species turnover along the axes (Hill and Gauch 1980, Gauch 1982*a*). Such artifices do eliminate the major problems with CA, but they introduce inelegancies that have been justly criticized for their uncertain effects (Minchin 1987*a*, Oksanen 1987, 1988, Wartenberg et al. 1987).

Canonical Correspondence Analysis.—Like DCA, the most common algorithm for CCA involves the addition of steps to CA (Fig. 1C). However, the new steps are added not to remove an undesirable effect, but to take advantage of supplemental data in the form of environmental variables. This is what makes CCA a *direct* gradient analysis. A multiple linear least-squares regression is performed with the site scores (determined from weighted averages of species) as the dependent variables, and the environmental variables as the independent variables. New site scores are now assigned as the value predicted using the regression equation. Since this regression equation is formally a Linear Combination of variables, let us label the new site scores LC scores, in contrast to the site scores determined by Weighted Averaging (WA).

Although the CCA solution is most commonly obtained by a weighted averaging algorithm, the solution is essentially an eigenanalysis, and can hence be obtained by any eigenanalysis algorithm (ter Braak 1986, 1987*c*). Indeed, Chessel et al. (1987) present a more efficient eigenanalysis solution for CCA. Nevertheless, the weighted averaging algorithm is sufficiently rapid and accurate for practical use, and is discussed here because of its historical importance and intuitive appeal.

The statistical model underlying CCA is that a species' abundance or frequency is a unimodal function of position along environmental gradients. CCA is an approximation to Gaussian Regression under a certain set of simplifying assumptions, and is robust to violations of those assumptions (ter Braak and Prentice 1988). CCA is inappropriate for extremely short gradients, in which species abundance or frequency is a linear or monotonic function of gradients (ter Braak 1987*b*, ter Braak and Prentice 1988). For further details on the nature of the statistical models underlying CCA and other members of the CA family, the reader is referred to Lebreton et al. (1990), Sabatier et al. (1989), ter Braak (1985, 1986, 1987*b–d*, 1988), and ter Braak and Looman (1986, 1987).

Since CCA, by any algorithm, produces two sets of site scores, it is unclear which is the most appropriate

TABLE 1. Parameters used in COMPAS simulations (unless otherwise stated). See Minchin (1987*b*) for computational details.

Two gradients
24 sites on a 6 × 4 regular grid
300 species
Maximum abundance for species lograndomly distributed from 1 to 100
Species ranges on both gradients taken from a normal distribution, $\mu = 100$, $\sigma = 30$
Species modes from uniform random distribution between −95 and 195
Alpha and gamma (skewness parameters) taken from uniform random distribution between 0.5 and 3.5
Quantitative noise taken from the normal distribution, and proportional to the square root of abundance

to use in an ordination diagram. The initial publications on CCA do not advise whether to plot WA scores or LC scores (ter Braak 1986, 1987*a–d*). Most papers using CCA fail to state which site scores are used. Even the manual for the program CANODRAW (Smilauer 1990) designed to plot CCA results does not state which set of scores is used, although a computer file accompanying the program indicates that LC scores are the default. The most recent version of CANOCO (the leading computer program for CCA) employs LC scores as the default, whereas previous versions utilized WA scores (ter Braak 1990). I suggest that ecologists use linear combinations in most cases, for reasons to be discussed below.

There is yet another variant of CA known as Detrended Canonical Correspondence Analysis (DCCA, ter Braak 1986, 1987*a*). As the name implies, DCCA incorporates *both* a detrending step and a linear regression step into the reciprocal averaging algorithm. I intend to argue that detrending is unnecessary for CCA.

The anatomy of CCA diagrams

Like CA and DCA, CCA allows the simultaneous plotting of species and site scores as points in an ordination diagram known as a *joint plot.* CCA has an additional benefit: environmental variables can be represented by arrows along with the species and site scores in a diagram known as a *triplot.* If the appropriate form of scaling is used (see ter Braak 1990), the length of an arrow indicates the importance of the environmental variable, the direction indicates how well the environment is correlated with the various species composition axes, the angle between arrows indicates correlations between variables, the location of site scores relative to arrows indicates the environmental characteristics of the sites, and the location of species scores relative to the arrows indicates the environmental preferences of each species.

Details of the interpretation of CCA diagrams are given in ter Braak (1986, 1987*a–d,* 1990), and excellent examples of such diagrams include Stevenson et al. (1989), Whittaker (1989), Wiegleb et al. (1989), Allen and Peet (1990), Borgegård (1990), Carleton (1990), John and Dale (1990), Odland et al. (1990), Prentice and Cramer (1990), Pyšek and Lepš (1991), and Retuerto and Carballeira (1991).

Elegance of CCA

Ter Braak (1986) reveals that the CCA algorithm is conceptually simple and algorithmically elegant, and nicely unites two distinct bodies of statistical techniques (i.e., weighted averaging techniques and multivariate regression techniques). There is no reason, from simply studying the algorithms, that CCA should *not* work (this may be why CCA, unlike most other ordination techniques, has not previously been tested by simulation).

Unfortunately, elegance in the past has been deceptive. Extremely elegant techniques such as Principal Components Analysis and Canonical Correlation Analysis perform very poorly on most ecological data (Gauch and Wentworth 1976, Gauch 1982*a,* Pielou 1984, Digby and Kempton 1987, Minchin 1987*a,* ter Braak 1987*b*). It is clear that elegance alone is insufficient reason for accepting a multivariate method.

In this paper, I examine the behavior of CCA with data sets whose properties are completely known—namely, simulated data sets. Furthermore, I test CCA's performance with high levels of noise. Since CCA is part of the correspondence analysis family, I also test whether the newer technique has inherited any defects possessed by its relatives.

METHODS

Simulation of species distributions

I simulated species distributions using COMPAS, a computer program written by Minchin (1987*b*). COMPAS simulates species abundance along gradients as a beta function, which allows species to have nonsymmetrical, or skewed distributions along environmental gradients. In this study the default parameters for COMPAS are used (Table 1). These values result in skewed species distributions, and have been used to criticize the performance of DCA (Minchin 1987*a*).

Sites are situated as a 6 × 4 regular grid along two major (hypothetical) environmental gradients (Figs. 2 and 3). The simulated data consist of the abundance of each species in each site. It must be stressed that this design does not represent a spatial grid, but merely a grid in "ecological space" (sensu Gauch 1982). This sampling scheme is used in all simulations below unless otherwise stated. A grid design may not be realistic, but it allows rapid visual evaluation of the performance of a technique (Gauch 1982*a,* Kenkel and Orlóci 1986, Bradfield and Kenkel 1987, Minchin 1987*a*). Although Minchin (1987*a*) found differences in simulation results between sites placed in a regular grid and sites placed randomly, I tested both options and detected

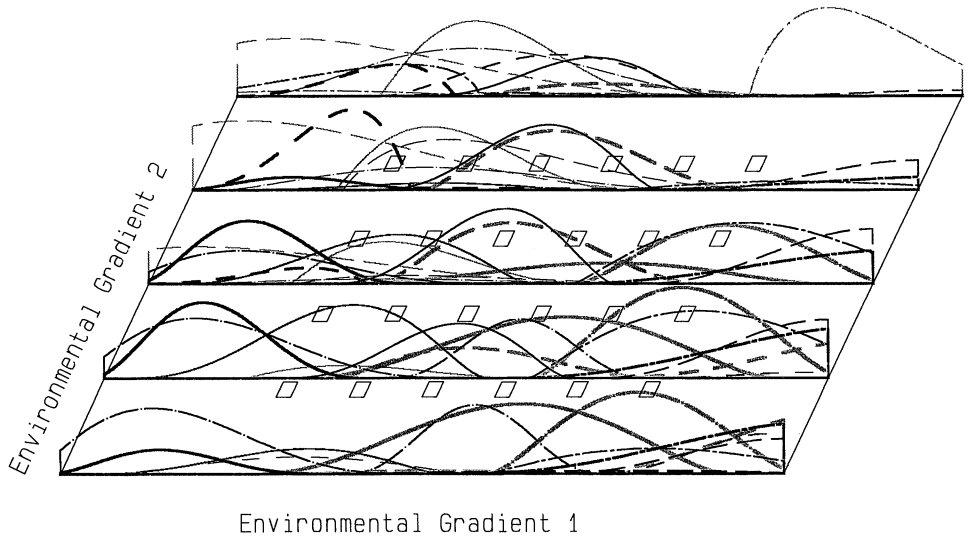

FIG. 2. The simulated distribution of the 26 most abundant species along the two simulated environmental gradients. The small rectangles indicate placement of sites along the gradients. The vertical axis indicates abundance (e.g., biomass) of species. Since it is virtually impossible to illustrate the abundance of many species simultaneously as a function of two gradients, "transects" are taken across the first environmental gradient at five different levels of the second environmental gradient (this is similar to the method employed by van Groenewoud 1992). The same species in different transects are indicated by different line styles.

no substantial differences; hence only the former are presented here.

Ordination

CCA and DCA were performed using the computer program CANOCO version 2.1 (ter Braak 1987*a*) with all the program defaults. One of the major choices made in DCA is whether to detrend by segments or by polynomials (ter Braak 1987*a*, Knox 1989, Økland 1990). Detrending is by polynomials in this paper. When detrending the simulated data by segments (not presented here) the configurations of the DCA diagrams were usually similar; however, when the two techniques produced dissimilar results the performance of both techniques was consistently poor.

An ideal ordination technique on the simulated data should result in a grid identical to that illustrated in Fig. 3. If CCA works optimally, there should be an arrow representing gradient 1 pointing to the right, and an arrow representing gradient 2 pointing perpendicular to it. Of course, no ordination technique will perform perfectly if there is an extremely high level of noise in the data. However, a robust and powerful technique should give results similar to those in Fig. 3 in spite of high noise.

If the only environmental gradients input into CANOCO were gradient 1 and gradient 2, we are practically guaranteed near-perfect results. This is because a regular grid will result as a linear combination of two perpendicular gradients. Unfortunately, we rarely know a priori what the most important gradients are. If we did, there would be little purpose in performing multivariate gradient analysis at all. We are usually more interested in determining which environmental variables represent real gradients and which variables are unimportant to species composition. In order to represent such "unimportant" variables, I input four variables in which the values were taken from a uniform random distribution, and which had no systematic relationships with simulated species abundance data. Thus the environmental data consist of two gradients and four random variables, labelled g1, g2, r1, r2, r3, and r4.

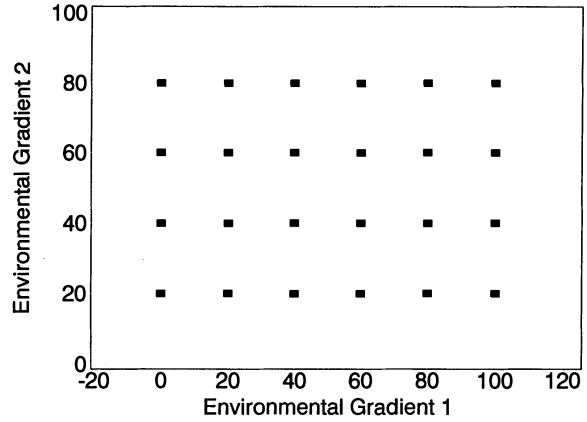

FIG. 3. The location of simulated sites along two simulated environmental gradients. This sampling design is used for the DCA and CCA analyses described below, unless otherwise stated.

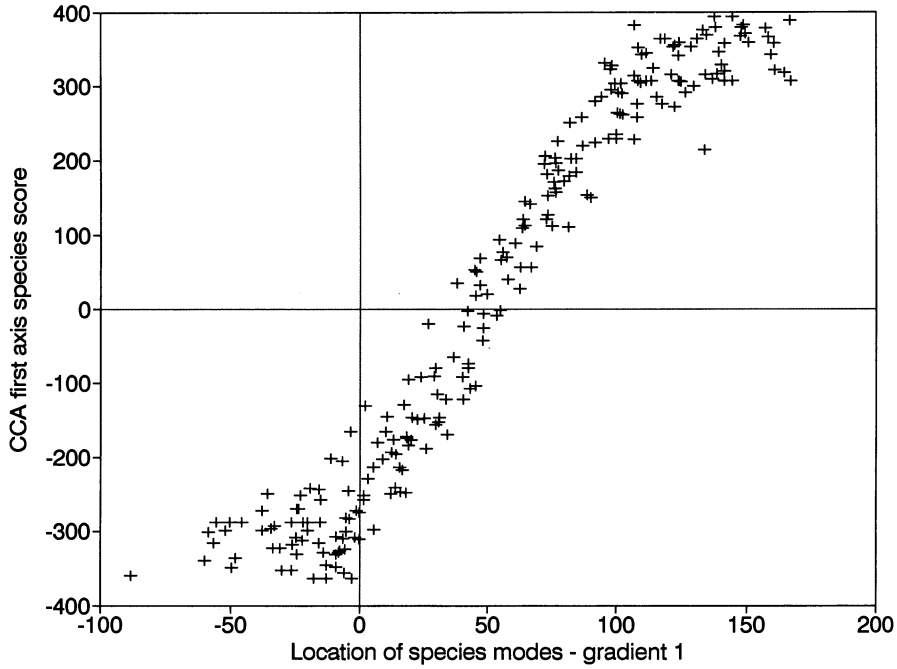

FIG. 4. Species scores for 300 simulated species along the first CCA axis, as a function of modal position along the simulated first gradient.

RESULTS

Species scores

Although most of this paper will concentrate on site scores, species scores will now be briefly considered. The perpendicular projection of species scores onto the environmental arrows are estimates of the modes of the species distributions. In fact, they are maximum likelihood estimates of species modes under the assumption that species abundance is a Gaussian function of environmental gradients (ter Braak 1986). Fig. 4 shows that species scores are good estimates of species modes even when species have highly skewed distributions. The relationship is poor when gradient positions are <0 or >100, but this is not at all surprising since there are no sites in these environments. No technique can be expected to adequately describe species responses outside of the range of sites.

Other deviations from a perfect relationship between modal positions and species scores can be attributed to two factors. First, the environmental arrow representing the dominant gradient is not exactly parallel to CCA axis 1 (Fig. 5). This will be discussed shortly. Second and more important, the modal position of a skewed curve is not identical to its weighted average position along a gradient. It can be argued that the weighted averaging position is a more valuable measure of position along a gradient than the mode, so the resulting scatter in Fig. 4 may be considered an appropriate result.

Site scores and environmental arrows

The CCA diagrams (Fig. 5) reveal that the grid of LC scores is clearly recovered with minimum distortion. Although the grid is slightly tilted, it *is* parallel with the environmental arrows for the two gradients, as desired (i.e., the configuration is similar to that in Fig. 3, if we take the arrows to be our axes). Note that the four arrows representing random gradients are quite short; as desired, they have almost no effect on the results.

In contrast to CCA, DCA warps the grid substantially for noiseless data. The warpage is either because the species distributions are skewed, or because of the tongue effect (Minchin 1987*a*, Økland 1990), or both. There are no environmental arrows because DCA is an indirect gradient analysis technique. Note that the WA scores from CCA are somewhat intermediate between the DCA results and the LC scores from CCA.

Quantitative noise

Of course, species abundance data typically possess much quantitative noise, and any multivariate technique would have little utility if it did not allow for this. There are several ways "noise" can be encountered in an ecological data set (Gauch 1982*a*, *b*, Lepš and Hadincová 1992). It can result from measurement error, an inadequate sampling intensity, or probably most important, stochastic variations of true abundance around the mean or ideal distribution.

Fig. 5 indicates a gradient of increasing noise in abundance from left to right. Noise here is presented in terms of a percentage of the square root of abundance for each species at each sampling location (Minchin 1987*b*). A noise level of 1000 is truly extreme; indeed, COMPAS does not allow higher values.

The LC scores of CCA are practically unchanged at even the highest noise levels. The length of the arrows for the random gradients increases, but minutely. DCA is barely affected by levels of noise from 0 to 100. However, a noise level of 1000 warps the grid substantially. Note in all cases how the CCA WA scores are intermediate between the DCA scores and the CCA LC scores.

Complex coenospaces

In this paper, "complex coenospaces" means that the sampling design is not well balanced along the major gradients. Unbalanced sampling designs can adversely affect the performance of DCA and other methods (Minchin 1987*a*). The results of some such sampling designs are illustrated in Fig. 6. These were produced without quantitative noise; when noise is added, the results are similar.

Minchin (1987*a*) demonstrated that DCA can distort the position of sites if the sampling design is T-shaped or cross-shaped. I did not find such extreme distortion in DCA (Fig. 6), but this may be because more species were simulated. The superior results in Fig. 6 are *not* due to the regular placement of sites along gradients; I obtained very similar results for randomly located sites within a T- or cross-shaped space (as was employed by Minchin 1987*a*).

Even though distortions by DCA are slight for the T- and cross-shaped designs, they are noticeable. CCA, however, has almost no distortion, and the arrows representing the real gradients are pointing in the correct directions. The four random gradients do have a noticeable but slight effect, by producing slight deviations in what should be straight lines.

One of the reasons for the development of DCA was the obliteration of the arch effect, which is usually a mathematical artifact (Hill and Gauch 1980, Gauch 1982*a*, Pielou 1984, Digby and Kempton 1987). One unfortunate consequence of this, however, is that DCA can destroy an arch even if it is a true property of the data. A true arch might exist, for example, if soils of circumneutral pH were invariably dry, whereas acidic and basic soils were always wet. Fig. 6 demonstrates that DCA does indeed destroy a true arch, whereas CCA preserves it, with the environmental arrows pointing in the correct directions. In all my simulations, this is the only circumstance in which I have observed an arch to appear in CCA (C. J. F. ter Braak [*personal communication*] suggests that an artificial arch may appear if a variable which is a quadratic function of the primary gradient is included in the environmental data. This is unlikely to occur in real data sets).

Fig. 6 also illustrates the situation where the first gradient dominates the second gradient. In this case, the grid is 12 × 2 rather than the 6 × 4 grid used in previous simulations. CCA is able to recover the second gradient, whereas DCA distorts it. Again, the CCA WA scores are intermediate between the CCA LC scores and the DCA scores.

Nonorthogonal and collinear gradients

"Nonorthogonal" and "collinear" are very similar concepts, but have different emphasis. By nonorthogonal, I mean that the most important gradients may be correlated with each other. By collinearity, I mean that there are a large number of variables included that are highly intercorrelated. Both of these factors have been considered problems in gradient analysis (Beals 1984, ter Braak and Looman 1987, Stergiou 1989).

In order to test how well CCA performs with nonorthogonal gradients, I created a new second gradient, which simply equals the value for the first gradient plus 0.01 times the value for the old second gradient. This creates two highly intercorrelated gradients, yet all the information about the second dimension is present in the environmental data. An ideal technique should be able to use this information. It can be argued that these gradients are so close that they don't "deserve" to be separated. Most ecologists, however, would prefer a technique that successfully reveals any meaningful relationships between species and environment.

Fig. 7 demonstrates that CCA can take advantage of subtleties in the environmental data. Although it appears that there is one arrow pointing to the right, in reality it is two arrows nearly on top of each other. The miniscule difference in the information contained in these two variables is entirely responsible for how well the entire grid is displayed: if the second gradient is not included in the analysis, as will shortly be described, the correct grid does not appear.

To simulate collinear gradients, I used six different environmental variables as input: three of them equal to the original first gradient plus a small random component (a uniform random number from 0 to 1) and the other three equal to the original second gradient plus a similar random component. Although the display for collinear gradients in Fig. 7 appears to be that of two more-or-less perpendicular arrows, there are in reality three arrows pointing in each direction. Thus creating collinear gradients does not "confuse" CCA into distorting the grid.

Gradients omitted from input

In direct gradient analysis, one is not always guaranteed that the most important environmental variables have actually been measured. A good test of a direct gradient analysis technique would be if the technique could still reveal relationships between environmental variables and species abundance, even if major

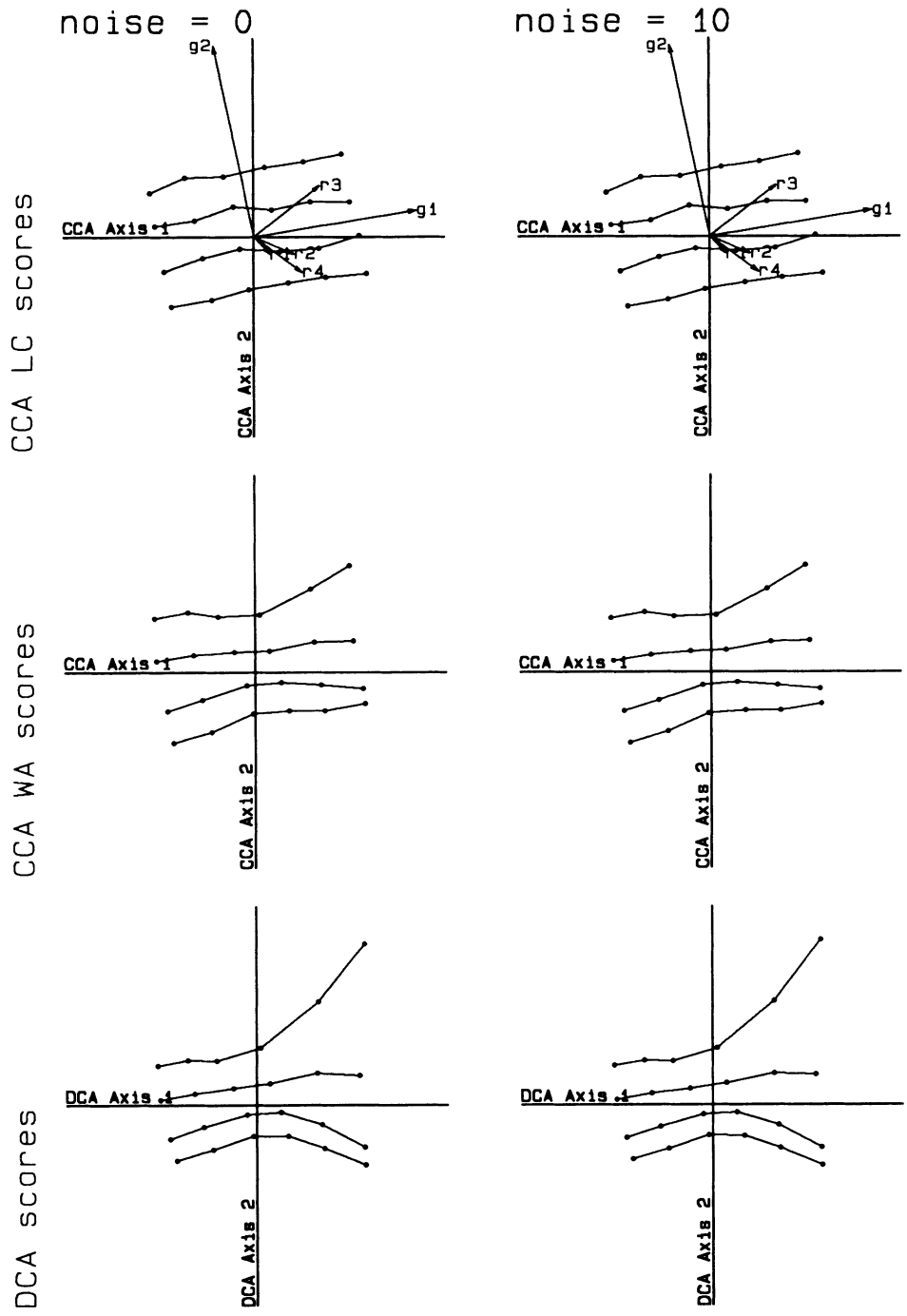

Fig. 5. Site scores along the first two axes in CCA and DCA ordinations, with varying levels of quantitative noise in species abundance. Qualitative noise was not simulated. The top set represents CCA LC scores and environmental arrows, the middle represents CCA WA scores, and the bottom represents DCA scores. Sites with equal positions along the environmental gradient 2 (see Fig. 3) are connected with lines to facilitate comparisons.

determinants of species composition were missing from the analysis.

Fig. 8 illustrates that when the first gradient is omitted from the analysis, the second gradient, as desired, is very close to parallel with the first CCA axis. If the second gradient is omitted, there is no trace of a grid along the second (or subsequent) CCA axes. Thus CCA tells us what the relationship is between the measured

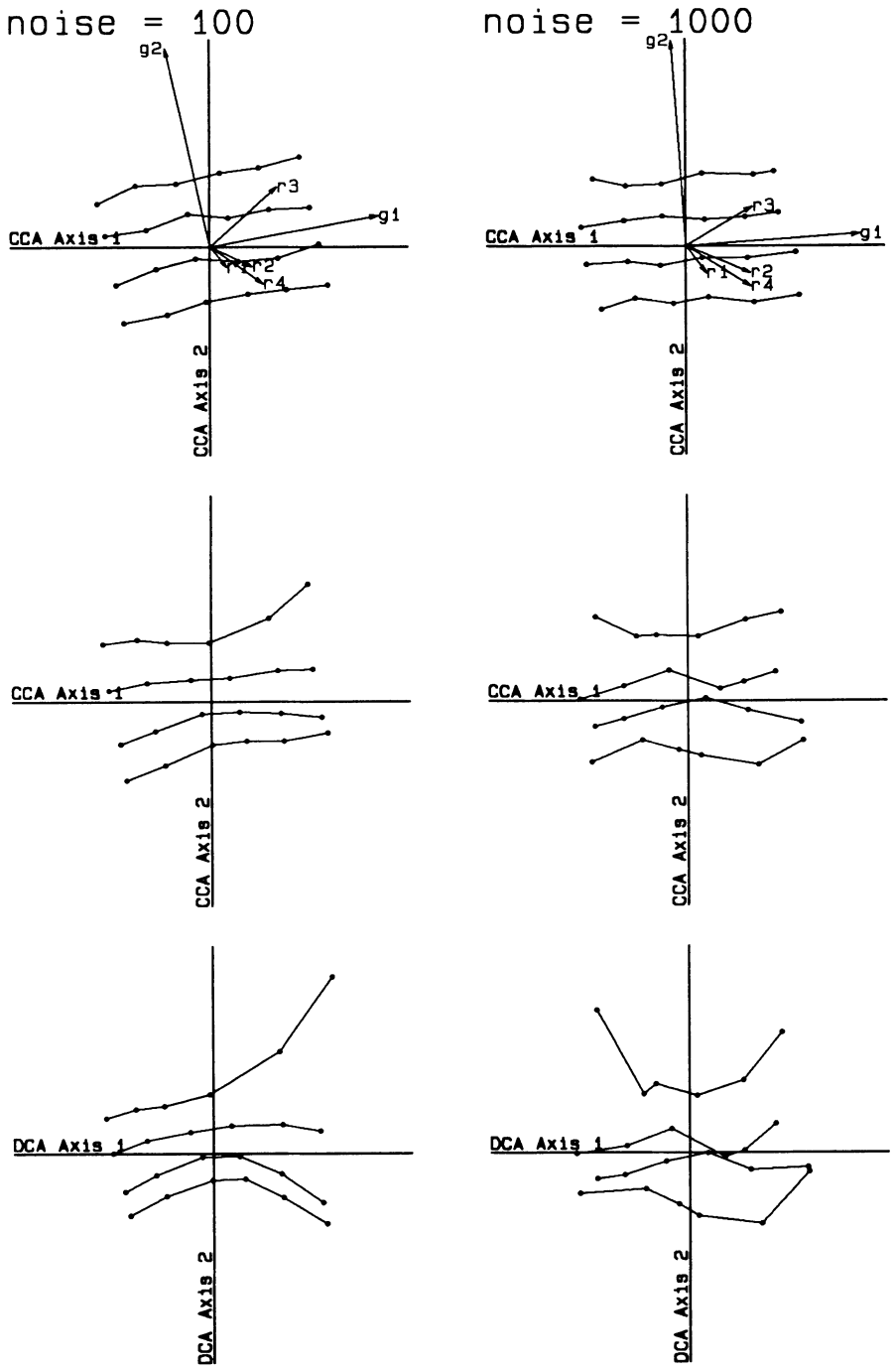

FIG. 5. Continued.

environmental variables and measured species composition, yet is not clouded by strong but unmeasured gradients.

Multidimensional coenospaces

Although I have only described the performance of CCA for two dominant gradients, I have found that CCA also performs well if there are three or four important gradients determining species composition. The desired result is no longer a two-dimensional grid, but rather a three- or four-dimensional regular array of points, which is difficult to display in a single figure. Although CCA performs as desired, I have not thoroughly tested performance with multidimensional

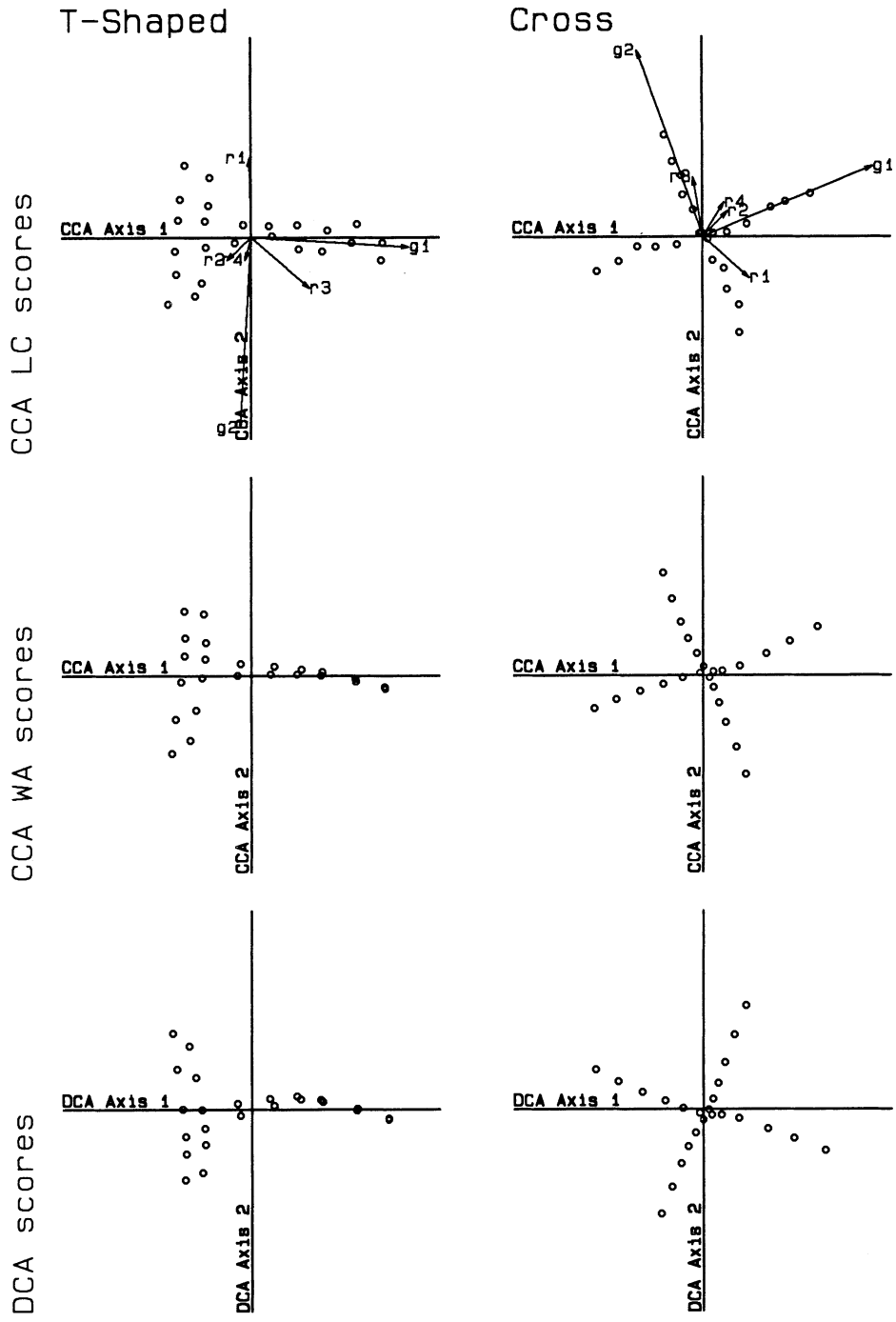

Fig. 6. Site scores along the first two axes in CCA and DCA ordinations, with different initial configurations of sites along the first two axes.

spaces using high noise levels, random site locations, or complex coenospaces.

Covariables

CCA offers a new opportunity for gradient analysis: the ability to "factor out" environmental variation in what is termed a *partial* ordination (ter Braak 1988). This could be very important, for example, if one wished to factor out site-to-site variation in testing for long-term successional patterns, or to factor out geological effects if the focus is on species responses to anthropogenic stress. The variables to be factored out are known as "covariables."

Whenever I have used covariables in simulated data

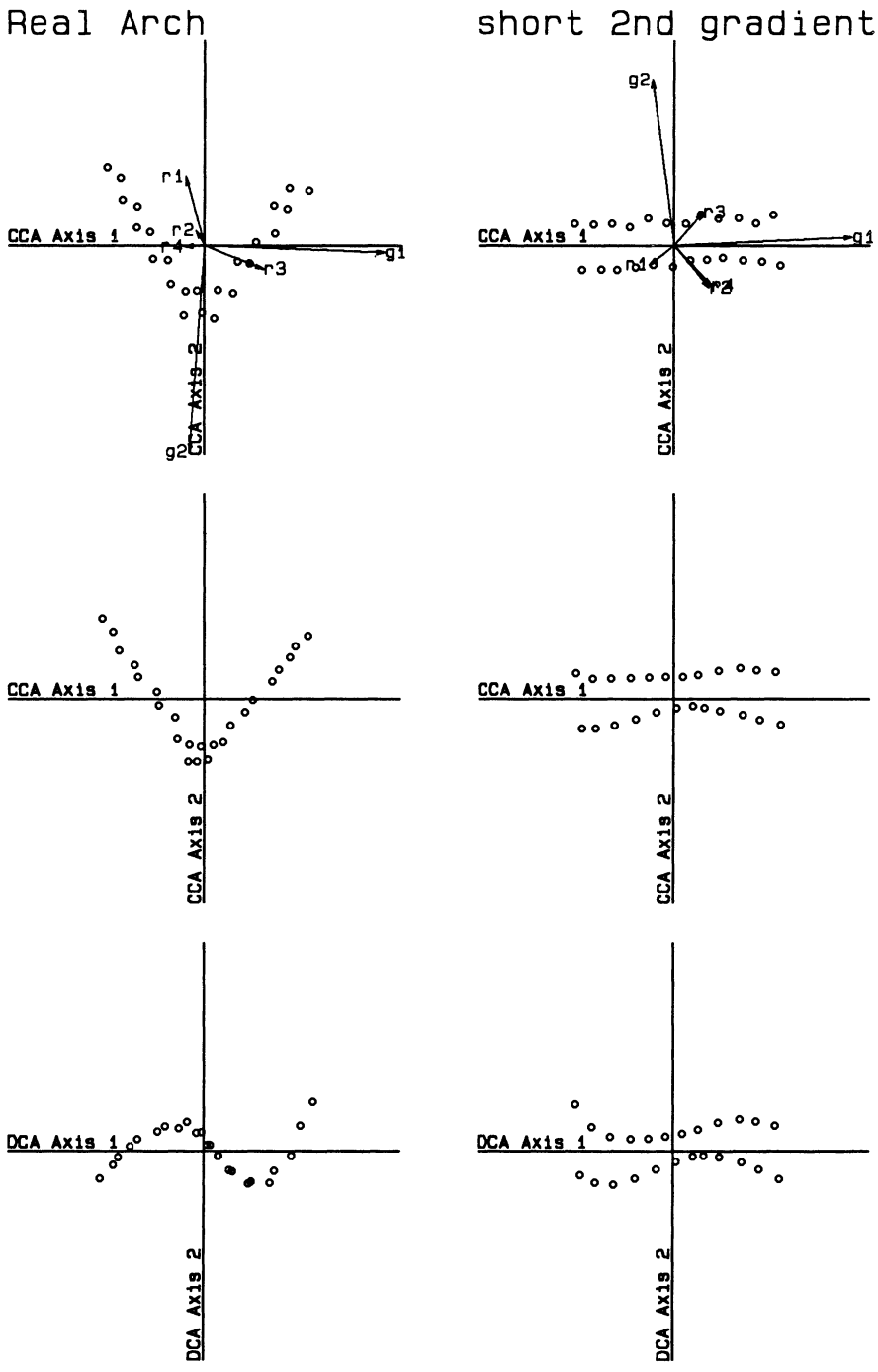

FIG. 6. Continued.

(that is, performed a partial ordination), I have found the desired result: there is no detectable trace of the covariable in the ordination diagram, and there is no major distortion of the grid. Again, I have not thoroughly tested this with high noise levels, random site locations, or complex coenospaces.

DISCUSSION

It may be argued that the simulations presented here are trivial. An examination of the CCA algorithm reveals that it should perform well. To this argument, I respond that seeing is believing. It is one thing to trust

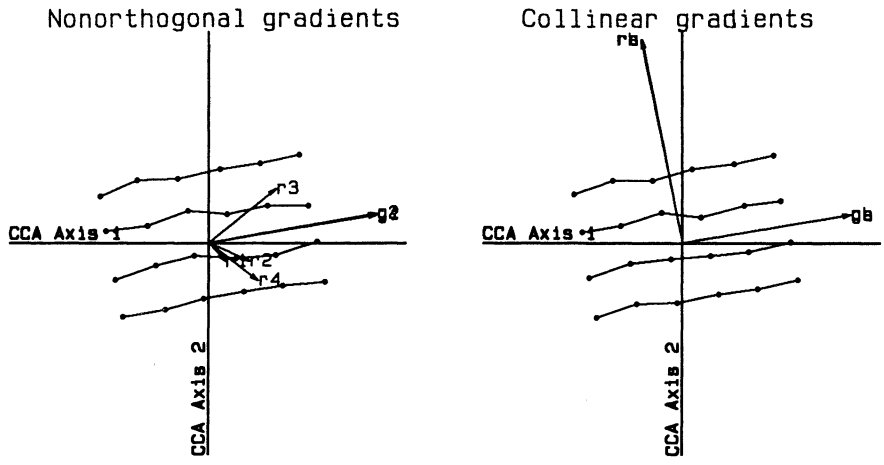

FIG. 7. Site scores along the first two CCA axes, when the environmental data that were input are nonorthogonal and collinear. The arrows for gradient 1 and gradient 2 are almost coincident in the diagram on the left. In the diagram on the right, three arrows representing environmental variables that are slight random deviations from gradient 1 are almost entirely coincident and pointing to the right; three arrows representing variables that are slight random deviations from gradient 2 are almost coincident and pointing upwards.

the validity of equations in the abstract, and yet another to entrust our data to them.

In general, CCA performs much better than DCA. However, DCA usually successfully uncovers the second ordination axis, albeit frequently with substantial warpage. This runs counter to the simulation results of van Groenewoud (1992), who concludes that correspondence analysis techniques fail to uncover axes beyond the first.

CCA performs well even if the data are not ideal. CCA performs well with skewed species distributions (Figs. 4 and 5) and extremely high noise levels (Fig. 5). It also performs well for complex sampling designs (Fig. 6). In addition, it will not generally create an artificial arch effect, but it will display an arch if it really exists. CCA does not display an undesirable "tongue effect," or compression of one of the gradient extremes (Minchin 1987*a*, Økland 1990). Thus CCA is immune to some of the defects of CA and DCA.

Since an artificial arch does not appear in CCA, detrending is not necessary. Detrending may even be harmful, because it may destroy a true arch or other complex sampling pattern. The only case in which Detrended Canonical Correspondence Analysis (DCCA) might be advisable is when detailed comparisons are made with DCA.

The ability of CCA to perform well with nonorthogonal and collinear gradients (Fig. 7) is reassuring because many environmental data sets consist of highly intercorrelated variables. For example, in the North Carolina piedmont, many variables (such as soil magnesium, calcium, cation exchange capacity, base saturation, etc.) are strongly correlated with soil pH (Christensen and Peet 1984, Palmer 1990).

One approach to such multicollinearity is to eliminate all of the variables but one. This approach is not always desirable. For example, it is possible that even if there is a strong positive correlation between calcium and magnesium, sites with high magnesium relative to calcium may still have distinct species compositions.

A second approach is to pre-process the environmental data by performing a multivariate analysis such as PCA, and choosing only the first several PCA axes as your environmental variables. This also is not desirable. For example, it is possible that a variable that contributes very little to the variance–covariance structure of the environmental data (and hence would be ignored in the analysis) actually has a strong influence on species composition. Another disadvantage of this approach is that the CCA diagram would become near-uninterpretable. For example, a CCA diagram with a long environmental PCA Axis III arrow parallel to the CCA Axis I would be nonsensical without a lengthy table of the PCA factor loadings for each environmental variable. Even with this table, it would be impossible to sort out which environmental variables are contributing to which species composition axes.

Fortunately, pre-processing of multicollinear data is unnecessary before using CCA. CCA can reveal a meaningful second axis even if the true variables are intercorrelated. This study demonstrates the truth of ter Braak's (1987*a*) statement, "The CCA ordination diagram is not in any way hampered by high correlations between species, or between environmental variables." Such redundancy in the environmental data is probably actually beneficial, because some errors in measuring the environmental data may be averaged out.

One advantage of CCA not tested in this study is that it is possible to test the significance of environ-

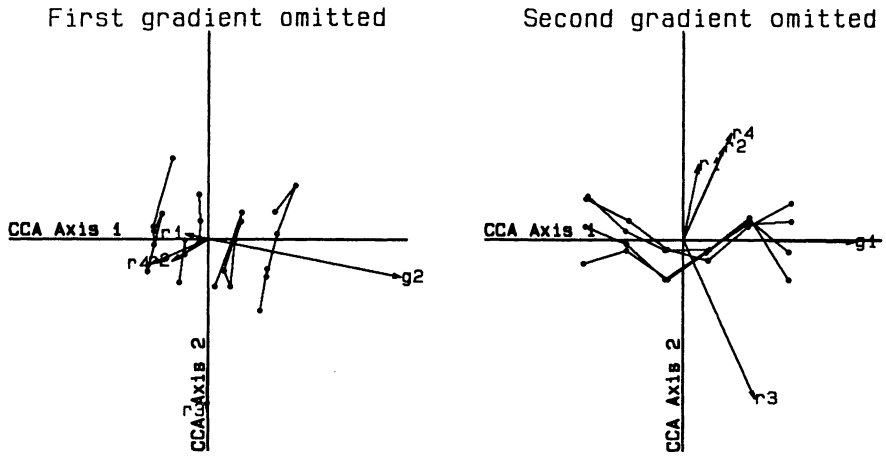

FIG. 8. Site scores along the first two CCA axes, when gradients are omitted from the environmental data set that was input.

mental variables using a Monte Carlo permutation test (ter Braak 1987*a*, *d*, 1988). Such tests are only valid if the sites are located objectively (preferably randomly) with respect to the environment, and sites are spatially independent. (However, version 3.10 of CANOCO does have the capability of factoring out some forms of spatial dependence.) If such criteria are met, Monte Carlo tests are quite appropriate as they do not make statistical assumptions concerning the distribution of environmental variables or species abundances. However, the statistical power of such Monte Carlo tests is difficult to ascertain.

In the vast majority of cases, CCA is likely to be used as an exploratory technique, based on sites that are subjectively located in what appears to be homogeneous ecological communities. If so, our goal is often to isolate a subset of environmental factors that leads to a reasonable interpretation of important gradients in a few dimensions. Although inferential statistics are no longer valid in this case, one could still use the regression capabilities of CCA to select those aspects of the environment that ideally explain variation in species composition. Recent versions of CANOCO allow one to perform Stepwise CCA, which is directly related to one of the mainstays of exploratory analysis, stepwise linear regression (Draper and Smith 1981). Although *P* values and other inferential statistics from Stepwise CCA are suspect, the end result is desirable: an ordination display with much lower dimensionality (and hence much higher interpretability) than the original data set. Stepwise CCA will include collinear variables if they have important contributions to variation in species composition, but it will pare down the number of completely redundant variables.

CCA presents us with two sets of site scores: the LC scores and the WA scores. This presents us with a dilemma: which is the most appropriate set to plot in an ordination diagram? At first glance, it appears that the WA scores are most appropriate. This is because the multivariate regression step in CCA (Fig. 1C) is a "fit" to the WA scores in much the same way as a linear regression is a "fit" to a scatter plot, and it is customary to show the actual data values in a scatter plot rather than just the predicted values (analogous to the LC scores).

Upon further consideration, however, the WA scores are not so appropriate. The LC scores can be considered the maximally constrained scores (i.e, constrained by the environmental variables), whereas the WA scores from pure CA can be considered minimally constrained. The WA scores from CCA inhabit a vaguely defined region between the two extremes; they are semiconstrained. As has been noted from the simulation results, the CCA WA scores are often intermediate between the LC score solution and the DCA solution, so it is unclear what the precise value of the WA scores is. To unduly anthropomorphize, the site scores are trying to break free from the constraints of the linear combinations, and approach the correspondence analysis result. Since the meaning of the WA scores is unclear, I strongly recommend the use of LC scores in CCA diagrams.

One major limitation of CCA is that the independent (environmental) variables are assumed to be measured without error, and to be constant within a site. This problem is not easily solved; indeed, error in the independent variables is a major problem for linear regression in general (Draper and Smith 1981). Within-site variation is a serious problem for direct gradient analysis in general (Palmer and Dixon 1990) and is therefore not a specific flaw of CCA.

As with linear regression, mathematical transformations of independent variables can have a profound effect on CCA. Fortunately, since tests of significance in CCA do not depend on parametric distributional assumptions, we do not need to concern ourselves with

transforming variables to conform to a normal (or any other) distribution. This allows us to choose transformations on a priori grounds. In many cases it is unclear what these a priori grounds should be, but I strongly suggest logarithmic transformations for soil chemical data.

Assume that soil calcium is an important determinant of plant species composition. If you do not transform soil calcium, we are assuming that a difference between 1 and 10 mg/kg calcium is of the same importance as the difference between 1001 mg/kg calcium and 1010 mg/kg calcium. This assumption is undoubtedly false: the former is likely to profoundly affect plant growth and species composition, while the latter will likely have negligible effect. On a logarithmic scale, however, the difference between 1 and 10 is on comparable terms with the difference between 100 and 1000 (that is, the differences between the logarithms of these numbers are equal). This is biologically much more reasonable. For example, plant growth is rarely a linear function of resource levels; more typically it is strongly concave-down (Tilman 1982). Such concave-down curves become more linear if the resource levels are logarithmically transformed. In the absence of physiological data on the nature of species responses to resource gradients, I strongly suggest that most resource gradients (e.g., photon flux, nutrient levels, rainfall, etc.) be logarithmically transformed prior to data analysis.

The problem of choosing an appropriate transformation for environmental variables is akin to the problem of skewed species distributions. Differences in transformations will not affect the relative positions of species along gradients, but it will affect the symmetry of the species response curves (Økland 1986). Since CCA performs well with skewed species distributions, it is likely that it will also perform well with a less-than-perfect transformation of environmental data.

To conclude, CCA is a direct gradient analysis technique that is an elegant extension of the indirect gradient analysis technique, Correspondence Analysis. CCA has all of the advantages and none of the disadvantages of DCA. The method estimates the modal locations of highly skewed species distributions quite well. It is robust to violations of assumptions. The arch effect only appears if there is a true arch in data; Detrending CCA is therefore unnecessary and may even be harmful. The ability to factor out covariables and to test for statistical significance further extends the utility of CCA.

ACKNOWLEDGMENTS

I thank R. Allen, R. Knox, and R. Peet for stimulating discussions on CCA, and H. J. B. Birks, K. Burnham, S. McAlister, J. Rotenberry, C. J. F. ter Braak, J. Thioulouse, T. Wentworth, and an anonymous reviewer for useful comments on the manuscript.

LITERATURE CITED

Allen, R. B., and R. K. Peet. 1990. Gradient analysis of forests of the Sangre de Cristo Range, Colorado. Canadian Journal of Botany **68**:193–201.

Allen, T. F. H. 1987. Hierarchical complexity in ecology: a noneuclidean conception of the data space. Vegetatio **69**: 17–25.

Austin, M. P. 1985. Continuum concept, ordination methods, and niche theory. Annual Review of Ecology and Systematics **16**:39–61.

Beals, E. W. 1984. Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. Advances in Ecological Research **14**:1–55.

Belbin, L. 1991. Semi-strong hybrid scaling, a new ordination algorithm. Journal of Vegetation Science **2**:491–496.

Birks, H. J. B., and H. A. Austin. 1992. An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986–1991. Botanical Institute, University of Bergen, Bergen, Norway.

Borgegård, S.-O. 1990. Vegetation development in abandoned gravel pits: effects of surrounding vegetation, substrate and regionality. Journal of Vegetation Science **1**:675–682.

Bradfield, G. E., and N. C. Kenkel. 1987. Nonlinear ordination using flexible shortest path adjustment of ecological distances. Ecology **68**:750–753.

Bray, J. R., and J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs **27**:325–349.

Carleton, T. J. 1990. Variation in terricolous bryophyte and macrolichen vegetation along primary gradients in Canadian boreal forests. Journal of Vegetation Science **1**:585–594.

Chessel, D., J. D. Lebreton, and N. Yoccoz. 1987. Propriétés de l'analyse canonique des correspondances: une illustration en hydrobiologie. Revue de Statistique Appliquée **35**: 55–72.

Christensen, N. L., and R. K. Peet. 1984. Convergence during secondary forest succession. Journal of Ecology **72**:25–36.

Digby, P. G. N., and R. A. Kempton. 1987. Population and community biology series: multivariate analysis of ecological communities. Chapman and Hall, London, England.

Draper, N. R., and H. Smith. 1981. Applied regression analysis. Second edition. Wiley, New York, New York, USA.

Escoufier, Y. 1987. The duality diagram: a means for better practical applications. Pages 139–156 *in* P. Legendre and L. Legendre, editors. Developments in numerical ecology. Springer Verlag, Berlin, Germany

Ezcurra, E. 1987. A comparison of reciprocal averaging and non-centered principal components analysis. Vegetatio **71**: 41–47.

Faith, D. P., and R. H. Norris. 1989. Correlation of environmental variables with patterns of distribution and abundance of common and rare freshwater macroinvertebrates. Biological Conservation **50**:77–98.

Gauch, H. G., Jr. 1982*a*. Multivariate analysis and community structure. Cambridge University Press, Cambridge, England.

———. 1982*b*. Noise reduction by eigenvalue ordinations. Ecology **63**:1643–1649.

Gauch, H. G., and T. R. Wentworth. 1976. Canonical correlation analysis as an ordination technique. Vegetatio **33**: 17–22.

Hatheway, W. H. 1971. Contingency-table analysis of rain forest vegetation. Pages 271–313 *in* G. P. Patil, E. C. Pielou, and W. E. Waters, editors. Statistical ecology. Volume 3. Pennsylvania State University Press, University Park, Pennsylvania, USA.

Hill, M. O. 1974. Correspondence analysis: a neglected multivariate method. Applied Statistics **23**:340–354.

Hill, M. O., and H. G. Gauch, Jr. 1980. Detrended Correspondence Analysis: an improved ordination technique. Vegetatio **42**:47–58.

John, E., and M. R. T. Dale. 1990. Environmental correlates of species distributions in a saxicolous lichen community. Journal of Vegetation Science **1**:385–392.

Kenkel, N. C., and L. Orlóci. 1986. Applying metric and nonmetric scaling to ecological studies: some new results. Ecology **67**:919–928.

Knox, R. G. 1989. Effects of detrending and rescaling on correspondence analysis: solution stability and accuracy. Vegetatio **83**:129–136.

Kruskal, J. B. 1964*a*. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**:1–27.

——. 1964*b*. Nonmetric multidimensional scaling: a numerical method. Psychometrika **29**:115–129.

Lebreton, J. D., R. Sabatier, G. Banco, and A. M. Bancou. 1991. Principal component and correspondence analysis with respect to instrumental variables: an overview of their role in studies of structure-activity and species-environment relationships. Pages 85–114 *in* J. Devillers and W. Karcher, editors. Applied multivariate analysis in SAR and environmental studies. Kluwer Academic, Dordrecht, The Netherlands.

Lepš, J., and V. Hadincová. 1992. How reliable are our vegetation analyses? Journal of Vegetation Science **3**:119–124.

Minchin, P. R. 1987*a*. An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio **69**:89–107.

———. 1987*b*. Simulation of multidimensional community patterns: towards a comprehensive model. Vegetatio **71**: 145–156.

Odland, A., H. J. B. Birks, and J. M. Line. 1990. Quantitative vegetation–environment relationships in west Norwegian tall-fern vegetation. Nordic Journal of Botany **10**: 511–533.

Økland, R. H. 1986. Rescaling of ecological gradients II: the effect of scale on symmetry of species response curves. Nordic Journal of Botany **6**:661–669.

———. 1990. A phytoecological study of the mire Northern Kisselbergmosen, SE Norway. II. Identification of gradients by detrended (canonical) correspondence analysis. Nordic Journal of Botany **10**:79–108.

Oksanen, J. 1987. Problems of joint display of species and site scores in correspondence analysis. Vegetatio **72**:51–57.

———. 1988. A note on the occasional instability of detrending in correspondence analysis. Vegetatio **74**:29–32.

Palmer, M. W. 1990. Spatial scale and patterns of species–environment relationships in hardwood forests of the North Carolina piedmont. Coenoses **5**:79–87.

Palmer, M. W., and P. M. Dixon. 1990. Small scale environmental variability and the analysis of species distributions along gradients. Journal of Vegetation Science **1**:57–65.

Peet, R. K., R. G. Knox, J. S. Case, and R. B. Allen. 1988. Putting things in order: the advantages of detrended correspondence analysis. American Naturalist **131**:924–934.

Pielou, E. C. 1984. The interpretation of ecological data: a primer on classification and ordination. Wiley, New York, New York, USA.

Prentice, H. C., and W. C. Cramer. 1990. The plant community as a niche bioassay: environmental correlates of local variation in *Gypsophila fastigiata*. Journal of Ecology **78**:313–325.

Pyšek, P., and J. Lepš. 1991. Response of a weed community to nitrogen fertilization: a multivariate analysis. Journal of Vegetation Science **2**:237–244.

Retuerto, R., and A. Carballeira. 1991. Defining phytoclimatic units in Galicia, Spain, by means of multivariate methods. Journal of Vegetation Science **2**:699–710.

Sabatier, R., J. D. Lebreton, and D. Chessel. 1989. Principal component analysis with instrumental variables as a tool for modelling composition data. Pages 341–352 *in* R. Coppi and S. Bolasco, editors. Multiway data analysis. Elsevier Science, North Holland, The Netherlands.

Sibson, R. 1972. Order invariant methods for data analysis. Journal of the Royal Statistical Society **34**:311–338.

Smilauer, P. 1990. CANODRAW: a companion program to CANOCO for publication-quality graphical output. Microcomputer Power, Ithaca, New York, USA.

Stergiou, K. I. 1989. A method to cope with collinearity of ecological data sets in community studies. Coenoses **4**:91–94.

Stevenson, A. C., H. J. B. Birks, R. J. Flower, and R. W. Battarbee. 1989. Diatom-based pH reconstruction of lake acidification using canonical correspondence analysis. Ambio **18**:228–233.

ter Braak, C. J. F. 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics **41**:859–873.

———. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology **67**:1167–1179.

———. 1987*a*. CANOCO—a FORTRAN program for community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis. Version 2.1. ITI-TNO, Wageningen, The Netherlands.

———. 1987*b*. Ordination. Pages 91–173 *in* R. H. Jongman, C. J. F. ter Braak, and O. F. R. van Tongeren, editors. Data analysis in community ecology. Pudoc, Wageningen, The Netherlands.

——. 1987*c*. The analysis of vegetation–environment relationships by canonical correspondence analysis. Vegetatio **69**:69–77.

———. 1987*d*. Unimodal models to relate species to environment. Agricultural Mathematics Group, Wageningen, The Netherlands.

———. 1988. Partial canonical correspondence analysis. Pages 551–558 *in* H. H. Bock, editor. Classification and related methods of data analysis. North-Holland, Amsterdam, The Netherlands.

———. 1990. Update notes: CANOCO version 3.10. Agricultural Mathematics Group, Wageningen, The Netherlands.

ter Braak, C. J. F., and L. G. Barendregt. 1986. Weighted averaging of species indicator values: its efficiency in environmental calibration. Mathematical Biosciences **78**:57–72.

ter Braak, C. J. F., and C. W. N. Looman. 1986. Weighted averaging, logistic regression and the Gaussian response model. Vegetatio **65**:3–11.

ter Braak, C. J. F., and C. W. N. Looman. 1987. Regression. Pages 29–77 *in* R. H. G. Jongman, C. J. F. ter Braak, and O. F. R. van Tongeren, editors. Data analysis in community and landscape ecology. Pudoc, Wageningen, The Netherlands.

ter Braak, C. J. F., and I. C. Prentice. 1988. A theory of gradient analysis. Advances in Ecological Research **18**:271–313.

Tilman, D. 1982. Resource competition and community structure. Princeton University Press, Princeton, New Jersey, USA.

van Groenewoud, H. 1992. The robustness of Correspondence Analysis, Detrended Correspondence Analysis, and

TWINSPAN Analysis. Journal of Vegetation Science **3**: 239–246.

Wartenberg, D., S. Ferson, and F. J. Rohlf. 1987. Putting things in order: a critique of detrended correspondence analysis. American Naturalist **129**:434–448.

Whittaker, R. J. 1989. The vegetation of the Storbreen Gletschervorfeld, Jotunheimen, Norway. III. Vegetation–environment relationships. Journal of Biogeography **16**:413–433.

Wiegleb, G., W. Herr, and D. Todeskino. 1989. Ten years of vegetation dynamics in two rivulets in Lower Saxony. Vegetatio **82**:163–178.