

## INFLUENCE OF NOISY ENVIRONMENTAL DATA ON CANONICAL CORRESPONDENCE ANALYSIS

BRUCE MCCUNE

*Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon 97331-2902 USA*

**Abstract.** Canonical Correspondence Analysis (CCA) is an increasingly popular method for multivariate analysis of ecological community data. It is, however, one of the most potentially misleading multivariate methods for community analysis. Inclusion of noisy or irrelevant environmental variables can distort the representation of gradients in community structure. These hazards are illustrated with simulated community data sets having a known, simple, underlying structure, then introducing different kinds and degrees of noise into the environmental data. Because of its sensitivity to even a modest amount of noise in the environmental data, CCA with site scores as linear combinations of environmental variables is inappropriate when the objective is to describe community structure. These problems can be avoided by using traditional indirect ordination methods, where pure community structure is expressed, without any constraint imposed by the environmental variables. CCA can be appropriate, however, when the objective is to describe how species respond to particular sets of observed environmental variables.

**Key words:** *canonical correspondence analysis; community analysis; gradient analysis; ordination; simulated coenoclines; simulated gradients.*

### INTRODUCTION

Canonical correspondence analysis (CCA; ter Braak 1986, 1994) is unusual among the ordination methods used in community analysis in that the ordination of the community data matrix (by reciprocal averaging; RA or CA) is constrained by a multiple regression on its relationships to environmental variables. Because CCA uses data on environment to structure the community analysis, CCA has been called a method for “direct gradient analysis” (ter Braak 1986). In contrast, performing an ordination on just the community data, then secondarily relating the ordination to the environmental variables, allows an expression of pure community gradients, followed by an independent assessment of the importance of the environmental variables.

CCA is best suited to community data sets where: (1) species responses are unimodal (hump-shaped), and (2) the important underlying environmental variables have been measured. According to ter Braak (1986, 1994), unimodal species responses to environment cause problems for methods assuming linear response curves (such as principal components analysis or redundancy analysis) but cause no problems for CCA.

The second condition results from the environmental matrix being used to constrain the ordination results.

CCA is currently one of the most popular ordination techniques in community ecology. Many ecologists use CCA as if it is yet another ordination technique, when in fact they differ in objectives (Økland 1996). CCA is easily misused because it is a relatively complex method and options in the software can strongly affect the meaning of results. Furthermore, the performance of the method has not been adequately explored and documented in the literature.

In particular, a fundamental but poorly understood characteristic of CCA is how it responds to noisy data. The environmental matrices used in community ecology often contain variables that are measured out of convenience rather than selecting the “best” variables from the organismal point of view. In many cases environmental variables are moderately noisy or worse. Because CCA explicitly uses these variables in extracting the most important community gradients, the influence of noisy environmental variables on CCA deserves close scrutiny.

This scrutiny has not been received. Palmer (1993) used simulated community data of known underlying structure and added variables containing random numbers to the environmental matrix. However, he left the

TABLE 1. Contents of the data matrices. Each matrix contained 100 rows (sites).

| Name     | Columns                                    | Source   |
|----------|--|--|
| TENXTEN  | 40 species                                 | Smooth Gaussian responses to a two-dimensional environmental gradient, simulated by COMPAS (Minchin 1987)              |
| TENXTEN2 | 2 environmental variables                  | $X, Y$ coordinates for position of each simulated sample unit on the underlying two-dimensional environmental gradient |
| NOISFULL | 99 random environmental variables          | Random number generator  |
| NOISE10  | 10 random environmental variables          | Random subset of NOISFULL  |
| NOISMOD  | 2 environmental variables with added noise | A small random number (mean = 0, variance = 17% of the total) is added to each cell of TENXTEN2                        |
| NOISBOTH | 12 environmental variables                 | Combined the 2 variables in NOISMOD with 10 variables from NOISFULL  |

noiseless environmental variables intact. In this case it is not surprising that CCA still managed to extract the major gradients successfully.

In this paper CCA is put to stronger, more realistic tests. In one case, a moderate amount of noise is added to the variables representing the two underlying gradients. In the second case, the environmental matrix is replaced with different numbers of variables containing random numbers. The random variables represent measured environmental variables that are irrelevant to community composition. The third case is perhaps most realistic, in that it combines moderate noise added to the most important underlying environmental variables, along with a number of additional variables composed of random numbers.

#### BASIC METHOD FOR CCA

The basic method for CCA has been clearly elaborated elsewhere (ter Braak 1986, 1994, Palmer 1993). A key point for this paper, however, is that two sets of site scores are produced.

Assume that data matrix  $\mathbf{Y}$  contains nonnegative abundances,  $y_{ij}$ , for  $i = 1$  to  $n$  sample units and  $j = 1$  to  $p$  species;  $y_{+j}$  and  $y_{i+}$  indicate species totals and sample unit (=site) totals respectively. The environmental matrix  $\mathbf{Z}$  contains values for  $n$  sites by  $q$  environmental variables.

At one step in the iterative algorithm, sites scores ( $x_i^*$ ) are calculated as weighted averages of species scores,  $u_j$ . The term formed by the eigenvalue,  $\lambda$ , and a user-selected constant,  $\alpha$ , is a scaling factor (ter Braak 1986):

$$x_i^* = \lambda^{\alpha-1} \sum_{j=1}^p y_{ij} u_j / y_{i+}$$

A different set of site scores ( $x_i$ ) is produced based on weighted least-squares multiple regression of  $x_i^*$  on the environmental variables ( $\mathbf{Z}$ ). Weights are  $y_{i+}$  in the diagonal of the otherwise empty matrix  $\mathbf{R}$ . The regression coefficients  $\mathbf{b}$  are calculated as

$$\mathbf{b} = (\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}\mathbf{x}^*.$$

New site scores ( $x_i$ ) are calculated as the fitted values from the preceding regression:

$$\mathbf{x} = \mathbf{z}\mathbf{b}.$$

These site scores are, therefore, linear combinations of environmental variables. These are "predicted" values produced by the regression equations built into CCA. Thus,  $\mathbf{x}^*$  and  $\mathbf{x}$  are derived such that the correlation between them is maximized, subject to the constraint that each is orthogonal to all previously extracted axes.

Following Palmer (1993), the site scores produced by the weighted averaging ( $x_i^*$ ) will be called the WA scores. The site scores produced as linear combinations ( $x_i$ ) of the environmental variables will be called the LC scores.

#### METHODS

Species responses to a two-dimensional environmental space were created using the program COMPAS (Minchin 1987). I experimented with several artificial data sets, but all gave qualitatively similar results, so only one is presented here. For comparison with other ordination methods I chose the same data set discussed by McCune (1994). The sampling pattern was a  $10 \times 10$  grid over the entire ecological space. The species responses were generated for 40 species without systematic trend, qualitative noise, or quantitative noise. Species response functions had the following characteristics: lograndom distribution of modal abundances, normal distribution of ranges on the gradients, uniform random distributions of modal coordinates on the gradients, and a mixture of skewed and symmetrical functions (see Minchin 1987). The particular sample units used below had an average richness of 10 species per sample unit (range 6–15). The  $\beta$  diversity (amount of change along a gradient; Wilson and Mohler 1983) for gradient 1 was 5.9 half changes (2.0 gleasons [amount that species importance changes along a gradient summed over all species]) and for gradient 2, 4.6 half changes (2.2 gleasons).

Two-dimensional ordinations of the simulated data

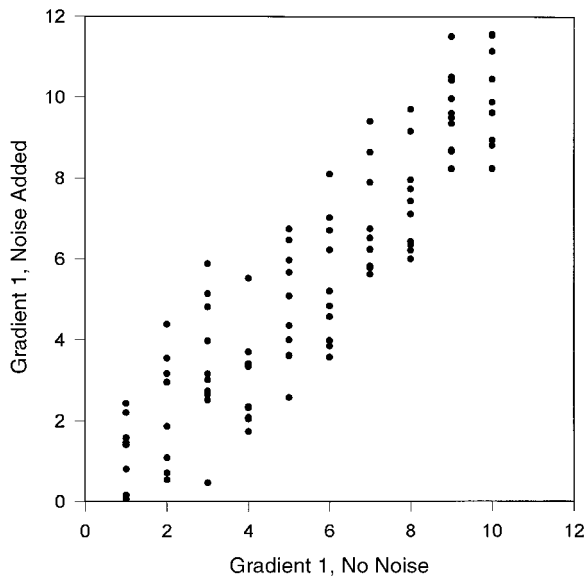


FIG. 1. Relationship between the original noiseless environmental variable representing a gradient and the variable transformed by adding noise.

sets (100 samples  $\times$  40 species) were produced. The environmental data consisted of grid coordinates expressing the  $x$ ,  $y$  position of each of the 100 sample units, analogous to positions on two independent environmental gradients. Then two kinds of noise were added to the environmental data, both singly and combined: (1) a modest random error added to the two environmental variables, and (2) replacement of the two environmental variables with from 1 to 99 random variables. In both cases random numbers were drawn from a uniform distribution. The data sets generated with these variables are listed in Table 1.

CCA in PC-ORD (McCune and Mefford 1995) was run with each of two alternative options for the site scores (LC scores and WA scores). CCA in PC-ORD has been tested against CANOCO (version 3.12; ter Braak 1990) to ensure consistent results. Quality of the ordination was assessed using the traditional criterion of minimal distortion of the underlying grid. For comparisons with other ordination methods using the same example data set, see McCune (1994). Although Procrustes analysis can be used to quantitatively compare ordination results, visual comparison of grid integrity provides a simpler, easily interpreted result.

#### *Influence of number of random variables*

An initial environmental matrix (NOISFULL; Table 1) contained 100 (sites)  $\times$  99 (random variables), each element of the matrix drawn from a random number generator with a uniform distribution between 0 and 1. The number of environmental variables was gradually

reduced by removing variables selected at random, performing a CCA at each step, and repeating this procedure until only one variable remained. At each step I recorded (for the first two axes) the species–environment correlation (the correlation coefficient between  $\mathbf{x}^*$  and  $\mathbf{x}$ ), the eigenvalue, and the percent of variance in the species data that was explained. These statistics are not affected by the choice of any of the axis re-scaling options in PC-ORD.

#### *Influence of adding noise to important environmental variables*

The two original environmental variables were made more realistic by adding a moderate amount of noise (NOISMOD; Table 1). The relationship between the transformed environmental gradient 1 and the original noiseless gradient 1 is shown in Fig. 1. Note that the relationship was fairly strong ( $r^2 = 0.83$  and the standard error about the regression line was 1.37). Thus, the transformed variables include a reasonably small amount of noise, a level that would be exceeded by many environmental measurements.

Ordinations were plotted using each kind of final site scores: those derived from the species data ( $\mathbf{x}^*$ , the WA scores) vs. those predicted by the regression equation ( $\mathbf{x}$ , the LC scores). In each case scores were standardized by centering and normalizing (biplot scaling), and scores were scaled to optimize the representation of chi-squared distances among sites. With this method, the scaling constant (alpha of ter Braak 1986) is set to 1 and site scores are weighted mean species scores.

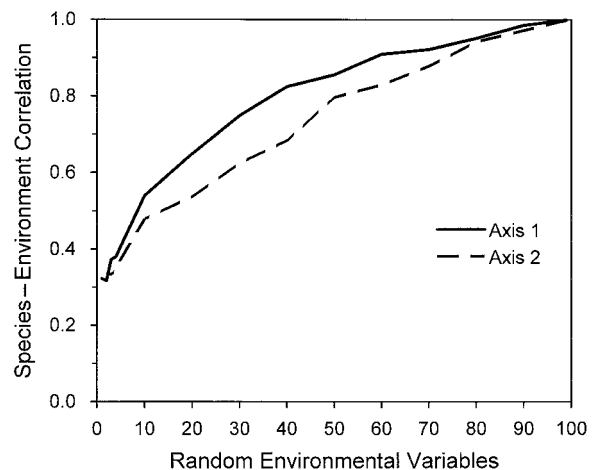


FIG. 2. Dependence of the “species–environment correlation,” the correlation between the LC and WA site scores, on a second matrix composed of from 1 to 99 random environmental variables. This correlation coefficient is inversely related to the degree of statistical constraint exerted by the environmental variables.

### Randomization tests

Statistical significance of eigenvalues and species–environment correlations was evaluated by randomization (Monte Carlo) tests, using 1000 randomized runs for each analysis. In each randomization, sample units in the environmental matrix were shuffled. This destroys the relationship between the species and environmental matrices, while preserving the species matrix and the correlation structure of the environmental matrix. The null hypothesis for a given parameter is that it is no larger than expected by chance. The probability of Type I error is calculated as  $(1 + n)/(1 + N)$  where  $n$  is the number of randomized runs in which the parameter from the real data was equalled or exceeded and  $N$  is the total number of randomized runs.

### RESULTS

#### *Influence of number of variables*

When the environmental matrix contains only one variable, and that variable is generated at random, the “species–environment” correlation is usually surprisingly high (0.32 in the example in Fig. 2), though statistically nonsignificant, for the first axis. This correlation rapidly increases as further random environmental variables are added until, when number of environmental variables is one less than the number of sample units, the species–environment correlation is 1.0 (Fig. 2). The same pattern is seen with the species–environment correlation on the second axis. In none of the cases, however, did randomization tests indicate statistical significance.

Proportion of variance in the species matrix that is explained has a similar pattern of increase, but the numbers are considerably lower and would not mislead the analyst into thinking a strong relationship existed. In this example, the first axis with a single random environmental variable explained only 1.6% of the variance in the species data, increasing to ~5% with ten random variables, and 19% with 99 random variables. The proportion of variance has, however, its own limitations, since its interpretation rests on a series of assumptions, e.g. that “inertia” (ter Braak 1990) adequately represents variation in communities and that ordination axes are not representing artifacts of curvature in gradients.

#### *Influence of adding noise to important environmental variables*

The influence of adding a small amount of noise (17% of the total environmental variation) to otherwise “good” environmental variables (Fig. 1) depends greatly on which set of site scores is used for the final ordination. If final site scores are linear combinations of environmental variables (LC scores), then the ideal representation with no noise (Fig. 3, top left) deteriorates as noise is added (Fig. 3, left column). The ideal representation consists of accurate reproduction of the underlying sampling grid.

In contrast, if final site scores are weighted averages from the species scores (WA scores), then the representation of the underlying sampling grid is never perfect, but it is insensitive to noisy environmental data (Fig. 3, column of graphs on right). In all cases the underlying sampling grid was recovered by the WA scores, but in a somewhat distorted way. In this case, having gradients of approximately equal length, the results are very similar to simple reciprocal averaging (=correspondence analysis) using only the species data.

When the ordination is based on the final LC scores, much of the deterioration in performance comes with the addition of a small error term to the two environmental variables representing the two “true” underlying environmental variables (compare the perfect noiseless representation in the top left graph of Fig. 3 with the next graph below it). When only random variables are used in the environmental matrix (NOIS10, Table 1), the underlying grid is, of course, completely distorted by the LC scores (Fig. 3). When the moderately noisy environmental variables are supplemented by 10 irrelevant variables (NOISBOTH, Table 1), the results are very similar to the inclusion of just the moderately noisy variables (Fig. 3, compare LC scores with NOISBOTH vs. NOISMOD). In other words, the inclusion of irrelevant variables has little effect on the results, provided that strong relationships are present with other environmental variables.

Randomization tests showed significant eigenvalues and species–environment correlations for all environmental matrices except those containing only random numbers (NOISE10 and NOISFULL; Tables 1 and 2).

→

Fig. 3. Influence of the type and amount of noise in the environmental data on CCA of simulated noiseless species data along two independent environmental gradients. Four different data sets represent different kinds and combinations of noise. Biplot vectors are shown for the noiseless environmental variables (top row) and with added noise (second and fourth rows). Row three contains biplot vectors for the three strongest random variables. Results are given for both kinds of site scores. Left column: LC site scores ( $\mathbf{x}$ ) are linear combinations of environmental variables. Right column: WA site scores ( $\mathbf{x}^*$ ) are derived by weighted averaging of species scores. “Species–environment correlations” ( $r$ ) are given at the right for Axis 1 and Axis 2, respectively.

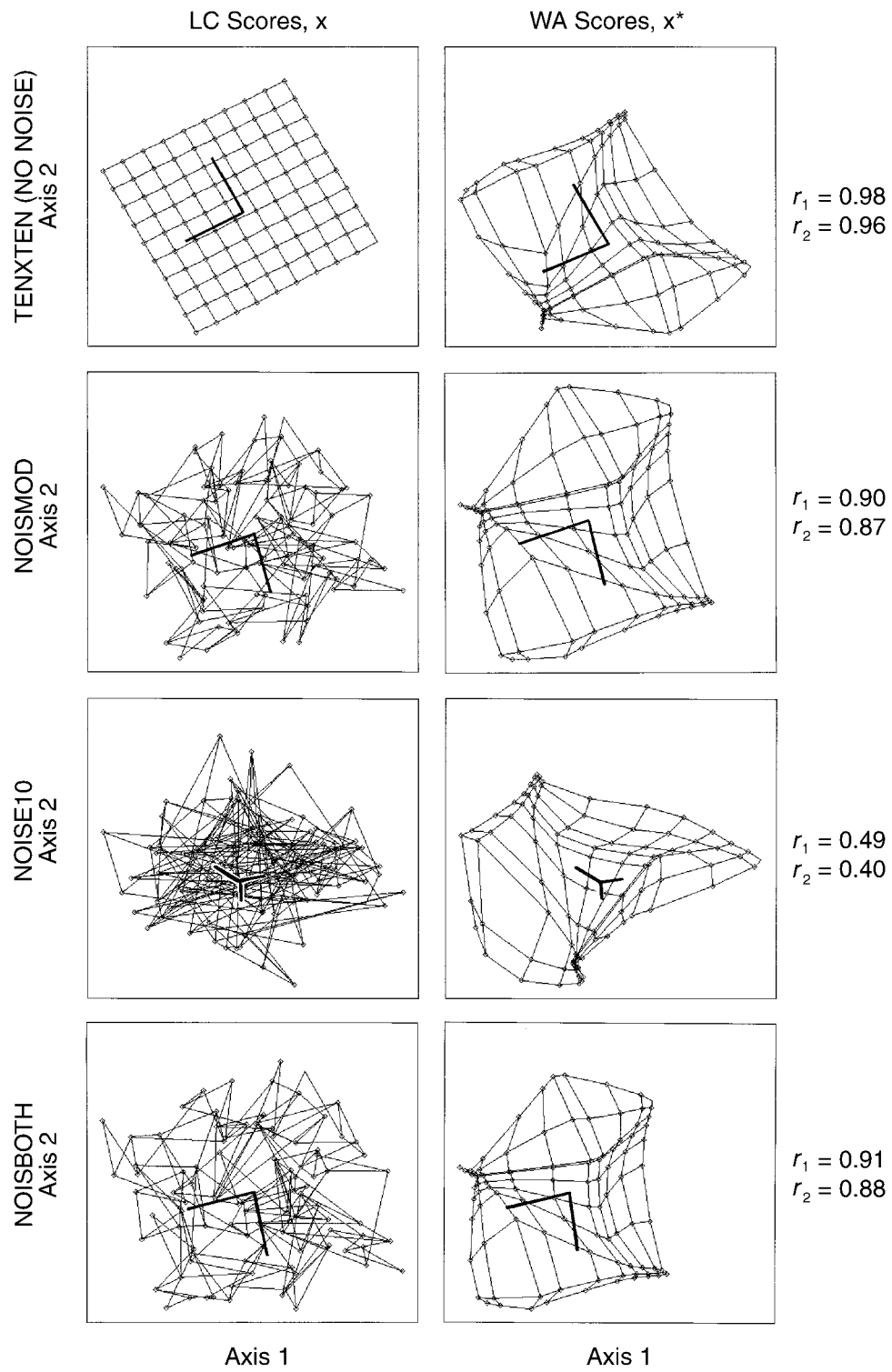


TABLE 2. Axis summary statistics for four environmental data sets representing different kinds and combinations of noise.

|                                 | Environmental data set |         |         |          |
|---------------------------------|------------------------|---------|---------|----------|
|                                 | TEXTEN2<br>No noise    | NOISMOD | NOISE10 | NOISBOTH |
| Eigenvalue                      |                        |         |         |          |
| Axis 1                          | 0.77*                  | 0.65*   | 0.20    | 0.65*    |
| Axis 2                          | 0.79*                  | 0.65*   | 0.12    | 0.66*    |
| Percent of variance             |                        |         |         |          |
| Axis 1                          | 17.5                   | 14.7    | 4.5     | 14.8     |
| Axis 2                          | 17.8                   | 14.7    | 2.7     | 15.0     |
| Species–environment correlation |                        |         |         |          |
| Axis 1                          | 0.98*                  | 0.90*   | 0.49    | 0.91*    |
| Axis 2                          | 0.96*                  | 0.87*   | 0.40    | 0.88*    |

Note: An asterisk (\*) indicates statistical significance based on randomization tests applied to eigenvalues and species–environment correlations.

Trials with other simulated data sets yielded results that were qualitatively the same.

#### DISCUSSION

##### *Meaning of the species–environment correlation*

The behavior of the “species–environment correlation,” when the environmental data are purely random, clearly indicates that this statistic has a misleading name. This correlation is simply the correlation between the LC site scores and the WA site scores. The LC site scores are controlled by the second (usually environmental) matrix, while the WA scores are directions of variation (community gradients) in the species matrix constrained to be maximally correlated with the LC scores.

As the number of variables in the second matrix increases to near the maximum possible (one less than the number of sample units), the species–environment correlation always converges on 1.0. At that point, the LC scores and WA scores are identical because the second matrix exerts no influence over the results (ter Braak 1994), since the large number of variables in the second matrix can support any pattern found by the weighted averaging step with the species matrix.

Using a wide variety of real data sets, it is clear that the species–environment correlation is almost always high. It is usually  $>0.6$ , and often much higher. This seems to be true regardless of other criteria for performance of the ordination, such as interpretability or proportion of variance in the species matrix that is explained.

For these reasons, I conclude that (1) the species–environment correlation is a poor criterion for evaluating the success of an ordination, (2) the species–environment correlation should not be interpreted literally as a measure of the strength of the relationship

between species and the environment, and (3) the statistical significance of the species–environment correlation, even when it appears very high, should always be checked with randomization tests.

The problems caused by large numbers of noisy environmental variables cannot be alleviated by using a stepwise selection procedure, similar to that commonly used in multiple regression. If the environmental data are noisy, stepwise selection will simply pick out the best random variables, and the species–environment correlation will still be misleadingly high. As with multiple regression, the parsimony of the procedure is set by the size of the pool of independent variables, not the number of independent variables selected.

I recommend that the species–environment correlation not be reported as such. If it is reported at all, a more appropriate name would be the “LC–WA correlation” to avoid misleading the reader. If it is reported, it should be accompanied by a randomization test for statistical significance. A better (but still imperfect) measure of the strength of the relationship between species and environmental matrices is the proportion of variance in the species data that is captured by the environmental data.

##### *Influence of adding noise to important environmental variables*

Even a modest amount of error in the environmental matrix can lead to serious distortions of the ordination space. Note that this is “distortion” only in the sense that the pure community gradients are destroyed; the ordination nevertheless accurately portrays the relationships between the species data and the noisy environmental data. These distortions appear only when final site scores are linear combinations of environmental variables (LC scores; i.e., true CCA). This was demonstrated by comparing the effects of three realistic alternative kinds of environmental noise to the unrealistic ideal of no noise in our environmental data.

If the objective was to describe community structure, then the distortion of the LC scores in the ordination space was unacceptably large in all three cases. CCA should not, therefore, be used for this goal when the environmental variables are measured with error.

##### *Choice of site scores*

Early versions of CANOCO (ter Braak 1988) and early papers on CCA used the WA scores. More recently ter Braak (1994) recommended the LC scores. This option is currently the default in both PC-ORD (McCune and Mefford 1995) and CANOCO (ter Braak 1990). The scores represent the best fit of species abundances to the environmental data. The WA scores best represent the community structure. The differences in

the resulting diagrams for the two sets of scores can be quite marked.

This choice has caused considerable confusion, as explained by Palmer (1993):

*Since CCA, by any algorithm, produces two sets of site scores, it is unclear which is the most appropriate to use in an ordination diagram. The initial publications on CCA do not advise whether to plot WA scores [derived from species matrix] or LC scores [linear combinations of environmental variables]. . . . Even the manual for CANODRAW . . . 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 important consequence for the choice of site scores is how the ordination reacts to noise in the environmental data. Scores which are linear combinations of environmental variables can be highly sensitive to noise in the environmental data. If the environmental data are irrelevant or noisy, then the resulting ordination is likely to mislead by showing spurious relationships. WA scores are insensitive to environmental noise, but they are not a direct gradient analysis in the sense that the axes are not combinations of environmental variables.

#### CONCLUSIONS

1) CCA is highly sensitive to noisy or irrelevant environmental data if one selects the option that final site scores are linear combinations of environmental variables. Use of LC scores will distort community patterns unless the data are noiseless. Because environmental data in community ecology usually contain measurement error, CCA with LC scores is inappropriate where the objective is to describe community structure (see further discussion of this point in Økland [1996]).

2) CCA can be appropriate, however, if the objective

is to describe community variation with respect to a particular set of measured environmental variables.

3) CCA is insensitive to noise if the final scores are derived from the species scores (WA scores), but in this case the axes are not combinations of environmental variables.

4) The “species–environment correlation” coefficient is a misnomer, is often misleadingly high, and if reported should be accompanied by a significance test. It would be better named the “LC–WA correlation.” A better measure of the strength of the species–environment relationship is the proportion of variance in the species matrix that is explained by the environmental matrix.

#### ACKNOWLEDGMENTS

I thank P. M. Dixon, P. Minchin, M. W. Palmer, C. J. F. ter Braak, and an anonymous reviewer for valuable suggestions for improving the manuscript.

#### LITERATURE CITED

- McCune, B. 1994. Improving community analysis with the Beals smoothing function. *Ecoscience* **1**:82–86.
- McCune, B., and M. J. Mefford. 1995. Multivariate analysis on the PC-ORD system. Version 2.0. MjM Software, Gleneden Beach, Oregon, USA.
- Minchin, P. R. 1987. Simulation of multidimensional community patterns: towards a comprehensive model. *Vegetatio* **71**:145–156.
- Økland, R. H. 1996. Are ordination and constrained ordination alternative or complementary strategies in general ecological studies? *Journal of Vegetation Science* **7**:289–292.
- Palmer, M. W. 1993. Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology* **74**:2215–2230.
- ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**:1167–1179.
- . 1988. CANOCO. Agricultural Mathematics Group. Technical Report LWA-88-02. Wageningen, The Netherlands.
- . 1990. Update Notes: CANOCO Version 3.10. Agricultural Mathematics Group. Wageningen, The Netherlands.
- . 1994. Canonical community ordination. Part I: Basic theory and linear methods. *Ecoscience* **1**:127–140.
- Wilson, M. V., and C. L. Mohler. 1983. Measuring compositional change along gradients. *Vegetatio* **54**:129–141.