Ecology, 83(4), 2002, pp. 1146–1161
© 2002 by the Ecological Society of America

NONLINEAR REDUNDANCY ANALYSIS AND CANONICAL CORRESPONDENCE ANALYSIS BASED ON POLYNOMIAL REGRESSION

VLADIMIR MAKARENKOV[1,2] AND PIERRE LEGENDRE[1,3]

[1]*Département de sciences biologiques, Université de Montréal, C.P. 6128, succ. Centre-ville,*
Montréal, Québec, Canada H3C 3J7
[2]*Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia*

Abstract. Among the various forms of canonical analysis available in the statistical literature, RDA (redundancy analysis) and CCA (canonical correspondence analysis) have become instruments of choice for ecological research because they recognize different roles for the explanatory and response data tables. Data table **Y** contains the response variables (e.g., species data) while data table **X** contains the explanatory variables. RDA is an extension of multiple linear regression; it uses a linear model of relationship between the variables in **X** and **Y**. In CCA, the response variables are chi-square transformed as the initial step, but the relationship between the transformed response data and the explanatory variables in **X** is still assumed to be linear. There is no special reason why nature should linearly relate changes in species assemblages to changes in environmental variables. When modeling ecological processes, to assume linearity is unrealistic in most instances and is only done because more appropriate methods of analysis are not available. We propose two empirical methods of canonical analysis based on polynomial regression to do away with the assumption of linearity in modeling the relationships between the variables in **X** and **Y**. They are called polynomial RDA and polynomial CCA, respectively, and may be viewed as alternatives to classical linear RDA and CCA. Because the analysis uses nonlinear functions of the explanatory variables, new ways of representing these variables in biplot diagrams have been developed. The use of these methods is demonstrated on real data sets and using simulations. In the examples, the new techniques produced a noticeable increase in the amount of variation of **Y** accounted for by the model, compared to standard linear RDA and CCA. Freeware to carry out the new analyses is available in ESA's Electronic Data Archive, *Ecological Archives*.

Key words: canonical correspondence analysis; multiple linear regression; nonlinear canonical analysis; permutation test; polynomial regression; redundancy analysis.

INTRODUCTION

Canonical analysis has become an instrument of choice for ecologists who want to relate a data table (**Y**) of response variables (which are often species abundances) to a second data table (**X**) of explanatory variables (often environmental factors). Two bibliographies of ecological papers on the subject, covering the periods 1983–1993 and 1994–1998 (H. J. B. Birks, S. M. Peglar, and H. A. Austin; and H. J. B. Birks, N. E. Indrevær, and C. Rygh, *unpublished manuscripts*), contain 804 titles. One can obtain a canonical ordination of the response variables whose axes are maximally and linearly related to the explanatory variables. Canonical analysis, which is also called constrained ordination analysis, provides interesting statistics, such as the proportion of variance of the response data that is accounted for by the explanatory variables, and tests of significance of this statistic and of individual canonical eigenvalues.

The forms of canonical analysis discussed in this paper are Redundancy Analysis (RDA) and Canonical Correspondence Analysis (CCA). Other forms of canonical analysis, such as canonical correlation analysis and discriminant analysis, are not of interest here. The development of Redundancy Analysis (RDA) is due to C. R. Rao (1964, 1973). RDA is an extension of multiple linear regression; it uses a linear model of relationships among the variables in **Y** and between the variables in **X** and **Y**. It may also be considered as a constrained extension of Principal Component Analysis (PCA) which identifies trends in the scatter of data points that are maximally and linearly related to a set of constraining (explanatory) variables. RDA consists of a series of multiple linear regressions followed by an eigenvalue decomposition of the table of fitted values. When table **Y** contains species abundance data, the component axes resulting from RDA are interpretable in terms of differences in the abundances of the species; thus the component axes in RDA biplots represent gradients in absolute species abundances, constrained by the explanatory variables.

Canonical Correspondence Analysis (CCA), developed by ter Braak (1986, 1987*a*) as an extension of

Manuscript received 9 May 2000; revised 17 April 2001; accepted 30 April 2001; final version received 4 June 2001.

[3] Please address correspondence to this author.
E-mail: pierre.legendre@umontreal.ca

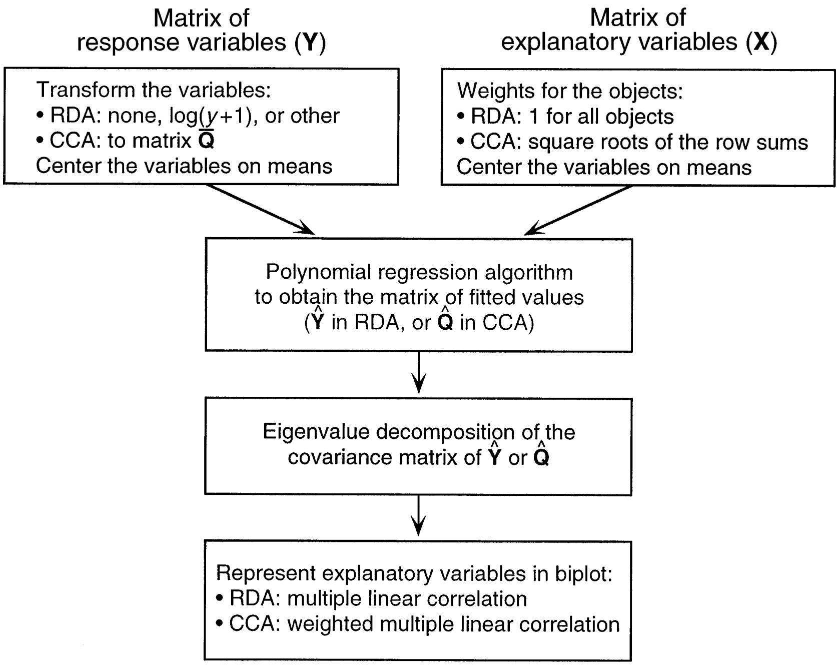
Matrix of response variables (Y)
Matrix of explanatory variables (X)
Transform the variables:
• RDA: none, log(y+1), or other
• CCA: to matrix $\bar{\mathbf{Q}}$
Center the variables on means
Weights for the objects:
• RDA: 1 for all objects
• CCA: square roots of the row sums
Center the variables on means
Polynomial regression algorithm to obtain the matrix of fitted values ($\hat{\mathbf{Y}}$ in RDA, or $\hat{\mathbf{Q}}$ in CCA)
Eigenvalue decomposition of the covariance matrix of $\hat{\mathbf{Y}}$ or $\hat{\mathbf{Q}}$
Represent explanatory variables in biplot:
• RDA: multiple linear correlation
• CCA: weighted multiple linear correlation

FIG. 1. RDA and CCA using polynomial regression.

Correspondence Analysis (CA), approximates unimodal responses of the species to environmental gradients, but it still assumes linearity of the relationships between the variables in **X** and **Y**. A chi-square transformation of the species abundances is done, but the relationship between the transformed response data and the explanatory variables in **X** is assumed to be linear. The component axes resulting from CCA basically represent gradients in species proportions, constrained by the explanatory variables.

There is no special reason why nature should linearly relate changes in species assemblages to changes in environmental variables. When modeling ecological processes, to assume linearity is unrealistic in most instances and is only done because more appropriate methods of analysis are not available.

This paper proposes a canonical (or constrained) ordination method based on polynomial regression to do away with the assumption of linearity in describing the relationships between the variables in **X** and **Y** (Fig. 1). This method, which builds upon the pioneering work of Rao and ter Braak, may be viewed as a nonlinear alternative to classical RDA and CCA. Our strategy is to apply polynomial regression, whose use is justified below, to describe the relationship between each response variable **y** of **Y** and the explanatory variables in **X**, in place of multiple linear regression. This approach may allow a noticeable increase in the explained variation of **Y**, compared to the linear model. The new approach often produces greater significance of the model than the linear approach; the significance of a canonical ordination model can be assessed using a permutation test.

The polynomial regression algorithm described in this paper allows modeling of polynomial relationships between the matrices of response and explanatory variables considered in RDA and CCA. The matrix of fitted values $\hat{\mathbf{Y}}$ used in the analysis is no longer a linear combination of the explanatory variables in **X**, but their polynomial combination. In this study, we only considered polynomials for which the degree of any particular explanatory variable included in any term of the polynomial is one or two. The regression algorithm proposed in this paper does not aim at providing an optimal polynomial with a fixed number of terms; it only tries to explain a portion of the variance, reflecting nonlinearities in the relationships, that cannot be accounted for by a linear regression model. Økland (1999) noted that species composition data rarely meet the assumptions of the species response models which are implicit in various methods of ordination and constrained ordination analysis. The nonlinear adjustments proposed in the present paper provide a way to enhance the fit of the model to the data in such cases.

The problem of expressing nonlinear relationships in canonical analysis has been investigated in the past. Van der Burg and de Leeuw (1983) used alternating least squares to find optimal nonlinear transformations of discrete data in canonical correlation analysis. Durand (1993) used additive spline transformations in RDA; Donovan (1998) also used spline transformations to express nonlinearities in RDA and CCA. These authors noted that the shapes of the transformations they obtained were generally not interpretable. We investigated the use of polynomial regression with the same objective in mind. Polynomials offer an elegant and easy way to obtain approximations of nonlinear relationships of unknown functional forms. The resulting equation is linear in its parameters, but the relationship between the response and explanatory variables may be linear or not; the linear equation is the simplest form of a polynomial function. Finally, a polynomial equation is an algebraic function and can be represented graphically. From the ecological point of view, polynomials represent a more flexible tool than linear models, which are embedded in them, to describe relationships between the response and explanatory variables. Product terms retained in polynomial equations represent combinations of variables having significant impact on the response data while significant second-order terms represent nonlinear relationships between explanatory and response variables.

A FORTRAN program was used to carry out the computations (Polynomial RdaCca; see the Supplement). After computing polynomial RDA or CCA, users of this program can also perform standard RDA and CCA based on multiple linear regression and assess the difference in explained variation between the two models, linear and polynomial, using a specially-designed permutation test.

This paper is organized as follows. (1) The new method of polynomial regression is first presented. (2) Classical RDA based on multiple linear regression is described, as well as its polynomial generalization. (3) CCA and its polynomial generalization are then presented, followed by (4) a discussion about ways of representing the explanatory variables in biplots and (5) tests of significance in polynomial canonical analysis. (6) To illustrate the new methods, a classical ecological data set containing nonlinear species–environment relationships is reanalyzed using polynomial RDA and CCA.

Polynomial Regression Algorithm

The algorithm described in this section aims at expressing each response variable **y** separately as a polynomial function of the explanatory variables most related to it. The variables should have already been transformed, if necessary, to insure homoscedasticity of the response variables. Reduction of the number of explanatory variables in the polynomial regression is necessary to avoid overfitting the response variables; in the linear case, overfitting occurs when a response variable is fitted using a number of explanatory variables larger than $(n - 1)$ where n is the number of observations. The polynomial algorithm proceeds by successively reducing the matrix of explanatory variables **X** while increasing the value of the coefficient of multiple determination R^2 for the response variable **y** under study. This reducing procedure is applied independently to each response variable **y**, corresponding to a column of the matrix of response variables **Y**. Let **y** be one of the response variables, associated with a vector of data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. The algorithm is comprised of four basic steps, described below, which are repeated $(m - 1)$ times as the matrix of explanatory variables **X** with m columns is reduced to a single vector.

1) **X** is a matrix of explanatory variables of order $(n \times m)$. The variables in **X** are centered on their respective means in order to reduce the collinearity between the linear and quadratic terms of the polynomial, calculated below. Binary explanatory variables, that may stand alone or may be used to code for multistate qualitative descriptors, may or may not be centered on their means; this is up to the user. The first step consists in regressing **y** on all variables in **X** following a classical least-squares multiple linear regression model. We find the vector of fitted values $\hat{\mathbf{y}}$ using vector **b** of the regression coefficients:

$$\hat{\mathbf{y}} = \mathbf{Xb} = \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y}. \tag{1}$$

2) The second step is to obtain the vector of residual values ($\mathbf{y}_{\mathrm{res}}$) from the multiple regression:

$$\mathbf{y}_{\mathrm{res}} = \mathbf{y} - \hat{\mathbf{y}}. \tag{2}$$

3) The task of the third step is to select the pair of variables in **X** that provides the best quadratic approximation of $\mathbf{y}_{\mathrm{res}}$. To accomplish this selection, for each pair of columns j and k of **X**, we compute a multiple linear regression of vector $\mathbf{y}_{\mathrm{res}}$ on matrix $\mathbf{X}^{jk}$ (where j and k are upper indices) containing variables x_j, x_k, x_jx_k, x_j^2, x_k^2 as columns, plus a column of 1's. For example, let $j = 1$ and $k = 2$; a quadratic polynomial regression of the vector of residuals $\mathbf{y}_{\mathrm{res}}$ (from Eq. 2) on variables $\mathbf{x}_1$ and $\mathbf{x}_2$ is obtained by

$$\hat{\mathbf{y}}^{12}_{\mathrm{res}} = \mathbf{X}^{12}\mathbf{c}^{12} \tag{3}$$

where $\mathbf{c}^{12}$ is the vector of regression coefficients for explanatory variables $j = 1$ and $k = 2$, and matrix $\mathbf{X}^{12}$ is constructed as follows:

$$\mathbf{X}^{12} = \begin{bmatrix} \mathbf{x}_{11} & \mathbf{x}_{21} & \mathbf{x}_{11}\mathbf{x}_{21} & \mathbf{x}_{11}^2 & \mathbf{x}_{21}^2 & 1 \\ \mathbf{x}_{12} & \mathbf{x}_{22} & \mathbf{x}_{12}\mathbf{x}_{22} & \mathbf{x}_{12}^2 & \mathbf{x}_{22}^2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_{1n} & \mathbf{x}_{2n} & \mathbf{x}_{1n}\mathbf{x}_{2n} & \mathbf{x}_{1n}^2 & \mathbf{x}_{2n}^2 & 1 \end{bmatrix}.$$

If $\mathbf{x}_1$ is a binary {0, 1} variable that has not been centered on its mean, the fourth column of $\mathbf{X}^{12}$ should not be included in this matrix; likewise for variable $\mathbf{x}_2$. The reason for this is that the square of a binary variable is equal to itself. The vector of regression coefficients $\mathbf{c}^{12}$ is computed using least squares, like vector **b** of Eq. 1. The coefficient of multiple determination $R^2(1, 2)$ is computed for this regression. The procedure is repeated for every pair (j, k) of columns of **X**. Each time, the coefficient of multiple determination $R^2(j, k)$ is computed. The pair (j, k) providing the largest coefficient of determination, $R^2(j, k)$, is retained; this pair will be used in step 4.

4) The two columns j and k selected in step 3 are

combined to form a new joint column t in $\mathbf{X}$, which replaces j and k for the remainder of the analysis. The following formula is used to compute the new combined variable t for each observation i ($i = 1, \ldots, n$):

$$x_{it} = x_{ij}b_j + x_{ik}b_k + \hat{y}^{jk}_{\mathrm{res},i} \tag{4}$$

where the coefficients b are those of Eq. 1. Thus, matrix $\mathbf{X}$ is reduced and now is comprised of one column (i.e., one variable) fewer than before. This new column combines the terms corresponding to the contributions of j and k to the linear regression of $\mathbf{y}$ on $\mathbf{X}$ (Eq. 1) as well as the fitted values of the regression of residual vector $\mathbf{y}_{\mathrm{res}}$ on matrix $\mathbf{X}^{12}$. Therefore, a new combined explanatory variable t is formed, containing the linear and quadratic contributions to the fitting of $\mathbf{y}$ by variables j and k.

The four steps above are repeated ($m - 1$) times as matrix $\mathbf{X}(n \times m)$ is transformed into a matrix $\mathbf{X}(n \times 1)$, which is a simple vector. To obtain the final vector $\hat{\mathbf{y}}$ to be used in the analysis in place of $\mathbf{y}$, we perform a simple linear regression of $\mathbf{y}$ on $\mathbf{X}(n \times 1)$. It should be clear that the vector of fitted values $\hat{\mathbf{y}}$ is now a polynomial function of the explanatory variables in the matrix $\mathbf{X}(n \times m)$ considered at the beginning of the regression procedure. This procedure also guarantees that every single variable of $\mathbf{X}$ is expressed by linear and quadratic terms in the reduced vector $\mathbf{X}(n \times 1)$.

We would not be able to control the maximum degree of any single variable in the polynomial if the quadratic form was used in matrix $\mathbf{X}^{jk}$ of Eq. 3 in each of the ($m - 1$) iterations. To make sure that the degree of each variable $\mathbf{X}$ is at most two in any single term of the polynomial, the following rule for composing matrix $\mathbf{X}^{jk}$ is applied for any pair of variables (j, k), starting from the second pass through the algorithm: if column j is already a combined variable obtained by Eq. 4, then its quadratic contribution (column x_j^2) should not be included in $\mathbf{X}^{jk}$. The same applies to variable k if it is a combined variable. Thus matrix $\mathbf{X}^{jk}$ may have from four to six columns, depending on the nature of the variables j and k.

The maximum degree of the polynomial is not bounded; it was not our objective to do so. Control is only exerted upon the highest degree, which is two, of any one variable in a monomial. In the most extreme case, one may end up with a polynomial of order m. Polynomials generated by this algorithm contain subsets of the terms from the following model:

$$\begin{aligned}\hat{y} = {} & b_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_m\mathbf{x}_m \\ & + b_{m+1}\mathbf{x}_1^2 + \cdots + b_{2m}\mathbf{x}_m^2 \\ & + b_{2m+1}\mathbf{x}_1\mathbf{x}_2 + \cdots + b\prod_i \mathbf{x}_i \prod_{j(j\neq i)} \mathbf{x}_j^2.\end{aligned}$$

The algorithm is used to determine which terms should be kept or deleted. This flexibility, as well as the huge range of shapes that the polynomial can fit, are among the advantages of this method. Distribution functions, which reflect species responses to the environmental factors found in real-life patterns, may vary a lot from species to species. Polynomial modeling provides a way of representing this diversity of responses. Admittedly, a polynomial model allows only an approximation and not an exact representation of the nonlinear relationships, whose functional forms are not known, but this approach is still far more efficient than the approximation by a straight line or a plane, as in classical linear regression.

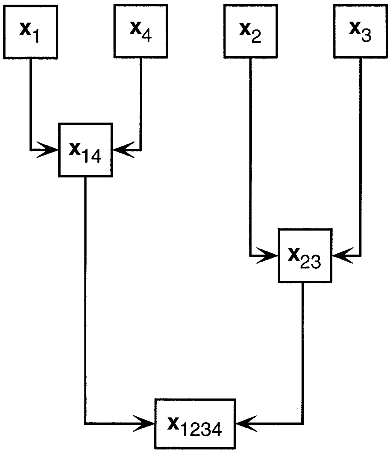

FIG. 2. Iterations of the polynomial weighted regression algorithm computed during the regression step of polynomial CCA for the first species (Sp. 1, *Al. accentuata*) in Table 1; the four explanatory variables are: $\mathbf{x}_1$, water content; $\mathbf{x}_2$, reflection of the soil surface; $\mathbf{x}_3$, percent cover by *Calamagrostis epigejos*; $\mathbf{x}_4$, percent cover by *Corynephorus canescens*.

Two examples of the use of this empirical procedure are presented in the section *Numerical examples*. Fig. 2 illustrates the computations for species 1 of the second example, where polynomial CCA is used. The detailed description of the polynomial regression results are presented in that section. The estimation of the number of independent parameters estimated by the polynomial regression procedure and the number of degrees of freedom left for statistical testing is described in Appendix A.

To obtain the matrix of fitted values $\hat{\mathbf{Y}}$ to be used in place of $\mathbf{Y}$ in the ordination analysis, ($m - 1$) passes through the algorithm are necessary for each response variable y_j ($j = 1, \ldots, p$) of $\mathbf{Y}$. Taking into account the $O(nm^2)$ time complexity of each loop consisting of the four steps described above, the whole algorithm performed on two matrices $\mathbf{Y}(n \times p)$ and $\mathbf{X}(n \times m)$ requires time on the order of pnm^3.

REDUNDANCY ANALYSIS AND ITS POLYNOMIAL GENERALIZATION

There are dedicated software packages available to perform classical RDA and CCA, such as CANOCO (ter Braak 1988*a*, *b*, 1990, ter Braak and Smilauer 1998) and RDACCA described in Legendre and Legendre

(1998:579). Although the algorithmic strategies used in these two packages differ, they lead to identical results. The approach of ter Braak is based upon the iterative application of averaging or weighted averaging equations; the ordination axes are computed one by one. In this work, we follow the direct computational approach described in Legendre and Legendre (1998). The main steps, implemented in the program RdaCca, are summarized in Appendix B. In the present section, we describe the modifications to that algorithm needed to obtain polynomial RDA.

Let **Y** be a matrix of response variables with *n* rows, representing the sites or objects, and *p* columns corresponding to the species or other variables under study. For instance, **Y** may be a matrix of the abundances of *p* species at *n* sites. Let **X** be a matrix of explanatory variables with *n* rows representing the same sites as in **Y** and *m* columns corresponding to the explanatory variables observed at these sites.

The objective of the polynomial regression algorithm, described in the previous section, is to explain a part of the variance of **Y** which remained unexplained after multiple linear regression. The approach is a direct modification of the algorithm for classical RDA presented in Appendix B. The first step is to calculate the polynomial regression of **Y** on **X**, i.e.,

$$\hat{\mathbf{Y}} = P(\mathbf{X}, \mathbf{X}^2) \tag{5}$$

where $P(\mathbf{X}, \mathbf{X}^2)$ denotes the polynomial equations of the previous section, which may differ for each variable **y** of **Y** not only in their polynomial coefficients but also in the **X** variables that are included in the equations. The covariance matrix **S** of $\hat{\mathbf{Y}}$ is computed in the classical way (Eq. B.2), followed by eigenanalysis of **S** (Eq. B.4). The site scores needed to represent the **Y** variables in biplots are calculated using equations of the same type as in principal component analysis (Eq. B.5 or B.6). In polynomial RDA, the matrix of eigenvectors **U** corresponding to non-null eigenvalues is of size $(p \times l)$ where l cannot exceed p or $(n - 1)$ but may be larger than *m*.

Each canonical ordination axis is now a quadratic function of the explanatory variables in **X**, the degree of each variable **X** in the polynomial being at most two. It is denoted as follows:

$$\mathbf{cord}_{(\text{space of explanatory variables } \mathbf{X})k} = \hat{\mathbf{Y}}\mathbf{u}_k = P(\mathbf{X}, \mathbf{X}^2)\mathbf{u}_k. \tag{6}$$

Canonical Correspondence Analysis and its Polynomial Generalization

We will now show how to use polynomial regression in the framework of canonical correspondence analysis (CCA). Basically, CCA is similar to RDA; the main difference is that it preserves chi-square distances, as in correspondence analysis (CA), instead of Euclidean distances among sites. Matrix $\hat{\mathbf{Q}}$ contains fitted values obtained by weighted linear regression of a matrix $\bar{\mathbf{Q}}$ of the contributions to chi-square (also used in CA) on the weighted explanatory variables found in matrix **X.** There are several algorithms for CCA. Appendix C outlines the one that served as the basis for this paper.

Let **Y** be a matrix of size $(n \times p)$ containing *p* species abundance or presence–absence variables, or other frequency data, observed at *n* sites. As in RDA, **X** is a matrix of explanatory variables of size $(n \times m)$, with rows representing the same sites as in **Y** and columns corresponding to the explanatory variables observed at the sites. The main differences from RDA are that **Y** will be chi-square transformed into a matrix $\bar{\mathbf{Q}}$, as in contingency table analysis, and that the rows of matrix **X** will be weighted by the square roots of the row sums of **Y**. An operational definition of matrix $\bar{\mathrm{Q}}$ is given in Appendix C.

In the present section, we describe how the algorithm for CCA outlined in Appendix C can be modified to incorporate the polynomial regression technique. As in RDA, the objective is to increase the percentage of variance accounted for, compared to standard CCA. In order to perform weighted polynomial regression, in place of weighted linear regression, in the first step of the analysis, we introduce the following changes to the equations of the polynomial regression procedure. Let $\bar{\mathbf{q}}$ be a variable corresponding to a single species from matrix $\bar{\mathbf{Q}}$ (Eq. C.1). To take weights into account, the changes to introduce into Eqs. 1–4 are the following:

$$\hat{\mathbf{q}} = \mathbf{X}_\mathrm{w}\mathbf{b} = \mathbf{X}_\mathrm{w}[\mathbf{X}'_\mathrm{w}\mathbf{X}_\mathrm{w}]^{-1}\mathbf{X}'_\mathrm{w}\bar{\mathbf{q}} \tag{7}$$

where $\mathbf{X}_\mathrm{w} = \mathbf{D}(p_{i+})^{1/2}\mathbf{X}$ is the weighted matrix of explanatory variables,

$$\bar{\mathbf{q}}_\mathrm{res} = \bar{\mathbf{q}} - \hat{\mathbf{q}} \tag{8}$$

$$\begin{aligned}\hat{\mathbf{q}}^{12}_\mathrm{res} &= \mathbf{D}(p_{i+})^{1/2}\mathbf{X}^{12}\mathbf{c}^{12}\\ &= \mathbf{D}(p_{i+})^{1/2}\mathbf{X}^{12}[\mathbf{X}^{12\prime}(\mathbf{D}(p_{i+})\mathbf{X}^{12}]^{-1}\\ &\quad \times \mathbf{X}^{12\prime}\mathbf{D}(p_{i+})^{1/2}\bar{\mathbf{q}}_\mathrm{res}.\end{aligned} \tag{9}$$

The following formula is used to compute the new combined variable *t* for each observation *i* ($i = 1, \ldots, n$):

$$x_{it} = x_{ij}b_j + x_{ik}b_k + p_i^{-1/2}\,\hat{q}^{jk}_{\mathrm{res},i} \tag{10}$$

Thus, a weighted polynomial relationship is described between matrix $\bar{\mathbf{Q}}$ of the contributions to chi-square and the matrix of explanatory variables **X**. After the polynomial regression procedure, one obtains:

$$\hat{\mathbf{Q}} = P_\mathrm{w}(\mathbf{X}, \mathbf{X}^2) \tag{11}$$

where $P_\mathrm{w}(\mathbf{X}, \mathbf{X}^2)$ denotes polynomials in which the highest degree of each variable is two, and whose coefficients depend on the weights. As in the case of polynomial RDA, the polynomial forms may vary from variable to variable.

The remainder of the analysis is based on matrix $\hat{\mathbf{Q}}$ and does not differ from the linear CCA outlined in Appendix C. The only remaining difference involves the computation of the scores of the explanatory variables **X** for biplots.

Representation of Explanatory Variables in Biplots

Two strategies can be used to represent the explanatory (e.g., environmental) variables in polynomial RDA and CCA biplots.

1) One can represent the individual terms of the polynomial by arrows in the biplot. When matrix **X** contains several variables, this strategy may produce too many terms (arrows) to be represented in the diagram. One may then apply an empirical rule, retaining only the correlations larger than a preselected value. In RDA, this strategy uses the correlations of the terms of the polynomial with the constrained ordination of the objects given by Eq. 6 and scaled using Eq. B.7. In CCA, the weighted correlations are obtained from Eq. C.12.

2) The second strategy is to represent each explanatory variable **x** by a single arrow in the biplot. This arrow corresponds to the multiple correlation of **x** and its quadratic form $\mathbf{x}^2$ with the ordination axes. It does not include any of the interaction terms. This solution may be preferred when there are so many explanatory variables in the analysis that the first strategy would produce a clogged diagram.

Let us examine this second option in more detail. In polynomial RDA, to obtain the biplot score of an explanatory variable **x** along a canonical ordination axis, a multiple correlation is computed between a vector of constrained ordination scores **cord** (from Eq. 6) and vectors **x** and $\mathbf{x}^2$, giving the multiple linear correlation $\boldsymbol{R}_{\mathbf{cord},\{\mathbf{x},\mathbf{x}^2\}}$. The sign of the simple linear correlation between **cord** and **x** is assigned to $\boldsymbol{R}_{\mathbf{cord},\{\mathbf{x},\mathbf{x}^2\}}$. With scaling type 2, $\boldsymbol{R}_{\mathbf{cord},\{\mathbf{x},\mathbf{x}^2\}}$ is used directly; with scaling 1, $\boldsymbol{R}_{\mathbf{cord},\{\mathbf{x},\mathbf{x}^2\}}$ is multiplied by coefficient c_k of Eq. B.7 to obtain the biplot score of **x** along the axis. Calculation of multiple linear correlations is described in Eqs. 12 and 13. For a binary {0, 1} variable **x** that has not been centered nor squared during polynomial regression, its score is obtained by simple linear correlation.

In polynomial CCA, a weighted multiple linear correlation is computed between the vector of constrained site scores **cord** (from Eq. C.10 or C.11) corresponding to a given axis and a pair of vectors $\{\mathbf{x}, \mathbf{x}^2\}$. The weights w_i in Eq. C.12, which are associated with rows i of vectors **cord**, **x**, $\mathbf{x}^2$, are equal to p_{i+}.

In polynomial RDA, let $\boldsymbol{R}_{\mathbf{cord},\mathbf{x}}$ be a coefficient of simple linear correlation; in polynomial CCA, it is a coefficient of weighted simple linear correlation between **cord** and **x**, computed using Eq. C.12. Let $\boldsymbol{R}_{\mathbf{cord},\mathbf{x}^2}$ and $\boldsymbol{R}_{\mathbf{x},\mathbf{x}^2}$ be the coefficients of (weighted) simple linear correlation between **cord** and $\mathbf{x}^2$, and between **x** and $\mathbf{x}^2$, respectively. The coefficient of (weighted) multiple linear correlation $\boldsymbol{R}_{\mathbf{cord},\{\mathbf{x},\mathbf{x}^2\}}$ is computed as follows:

$$R_{\mathbf{cord},\{\mathbf{x},\mathbf{x}^2\}} = \sqrt{1 - \frac{|\mathbf{M}_3|}{|\mathbf{M}_2|}} \qquad (12)$$

where matrices $\mathbf{M}_3$ and $\mathbf{M}_2$ are

$$\mathbf{M}_3 = \begin{bmatrix} 1 & R_{\mathbf{cord},\mathbf{x}} & R_{\mathbf{cord},\mathbf{x}^2} \\ R_{\mathbf{cord},\mathbf{x}} & 1 & R_{\mathbf{x},\mathbf{x}^2} \\ R_{\mathbf{cord},\mathbf{x}^2} & R_{\mathbf{x},\mathbf{x}^2} & 1 \end{bmatrix}$$

$$\mathbf{M}_2 = \begin{bmatrix} 1 & R_{\mathbf{x},\mathbf{x}^2} \\ R_{\mathbf{x},\mathbf{x}^2} & 1 \end{bmatrix}. \qquad (13)$$

The sign of $\boldsymbol{R}_{\mathbf{cord},\mathbf{x}}$ is assigned to $\boldsymbol{R}_{\mathbf{cord},\{\mathbf{x},\mathbf{x}^2\}}$ to obtain the biplot score of **x** along a canonical ordination axis corresponding to constrained ordination vector **cord**.

Biplot scores of the centroids of binary explanatory variables **x** are computed as in classical linear RDA and CCA. If w_i is the weight associated with row i of vectors **cord** and **x**, the score of the centroid of a binary variable **x** along vector **cord** is the following:

$$\text{Centroid}(\mathbf{x}, \mathbf{cord}) = \frac{\sum_{i=1}^{n} w_i cord_i x_i}{\sum_{i=1}^{n} w_i x_i}. \qquad (14)$$

Weights w_i are 1 in RDA and polynomial RDA. The arrow drawn using the biplot scores of a binary explanatory variable points toward this centroid, as in standard RDA and CCA based on (weighted) linear regression.

The biplot scores of the explanatory variables from matrix **X** are approximations of their real contributions in the full-dimensional space of canonical ordination. This point can be found in descriptions of biplots in Gabriel (1982), ter Braak (1994), and Legendre and Legendre (1998).

Tests of Significance in Polynomial RDA and CCA

Tests of significance can be carried out in linear or polynomial RDA or CCA. The most general null hypothesis is the same as in regression analysis; it states that there is no special relationship between the response and explanatory variables (independence of **Y** and **X**), or that the model is not a significant representation of the response data. The pseudo-F statistic used in the test as well as the method of permutation testing are described in Appendix D.

If the linear and polynomial models are both significant, another interesting question can be addressed: Which of the two models is the most appropriate to describe the data? To answer this question, a permutation procedure is used to assess the difference in variance accounted for, between the polynomial model and the linear model nested into it. Details of the method are presented in Appendix D.

Appendix D also reports the results of simulation studies showing (1) that our permutation test of sig-

TABLE 1. The number of individuals of hunting spiders caught in 28 traps (sites) over a period of 60 weeks, plus the values of four environmental variables measured at the same sites, from van der Aart and Smeenk-Enserink (1975).

Site no.	Sp. 1, *Al. accent.*	Sp. 2, *Al. cuneata*	Sp. 3, *Al. fabrilis*	Sp. 4, *Ar. lutetiana*	Sp. 5, *Ar. perita*	Sp. 6, *Au. albimana*	Sp. 7, *Pa. lugubris*	Sp. 8, *Pa. monticola*
1	25	10	0	0	0	4	0	60
2	0	2	0	0	0	30	1	1
3	15	20	2	2	0	9	1	29
4	2	6	0	1	0	24	1	7
5	1	20	0	2	0	9	1	2
6	0	6	0	6	0	6	0	11
7	2	7	0	12	0	16	1	30
8	0	11	0	0	0	7	55	2
9	1	1	0	0	0	0	0	26
10	3	0	1	0	0	0	0	22
11	15	1	2	0	0	1	0	95
12	16	13	0	0	0	0	0	96
13	3	43	1	2	0	18	1	24
14	0	2	0	1	0	4	3	14
15	0	0	0	0	0	0	6	0
16	0	3	0	0	0	0	6	0
17	0	0	0	0	0	0	2	0
18	0	1	0	0	0	0	5	0
19	0	1	0	0	0	0	12	0
20	0	2	0	0	0	0	13	0
21	0	1	0	0	0	0	16	1
22	7	0	16	0	4	0	0	2
23	17	0	15	0	7	0	2	6
24	11	0	20	0	5	0	0	3
25	9	1	9	0	0	2	1	11
26	3	0	6	0	18	0	0	0
27	29	0	11	0	4	0	0	1
28	15	0	14	0	1	0	0	6

Notes: The 12 species (Sp. 1 to Sp. 12) form the matrix of response variables **Y**. In the first example (RDA), the matrix of explanatory variables **X** contains the first two environmental variables, while in the second example (CCA), it contains all four. Water content is expressed as percentage of dry mass; reflection refers to reflection of soil surface under a cloudless sky (×100); *Calamagrostis* coverage is percent cover by *Calamagrostis epigejos*; *Corynephorus* coverage is percent cover by *Corynephorus canescens*.

nificance for polynomial RDA and CCA has correct type I error, and (2) that the test for the difference in explained variation between the polynomial and linear models also has correct type I error.

NUMERICAL EXAMPLES

Numerical examples of polynomial RDA and CCA are now presented. We used a well-known data set consisting of the abundance of 12 hunting spiders at 28 sampling sites, as well as the values of four environmental variables measured at the same sites. The data, displayed in Table 1, are from van der Aart and Smeenk-Enserink (1975: Tables 2 and 4) who studied them using PCA and canonical correlation analysis (CCoA). This data set has been reanalyzed by ter Braak (1986) in the paper where CCA was first described and by other authors since then. It contains several nonlinear species–environment relationships; examples are displayed in Fig. 3. This property was discussed by van der Aart and Smeenk-Enserink in their paper (1975). The polynomial equations for the relationships between the 12 species and two of the environmental variables are shown in Table 2. This data set is then ideally suited to display the advantages of polynomial canonical analysis.

Preliminary PCA of the log-transformed spider abundance data and CCA of the raw data confirmed the existence of a natural gradient in the data. The PCA ordination (not presented here; a similar ordination, including the same 28 plus 72 other traps, was published by van der Aart and Smeenk-Enserink [1975: Fig. 3]), had the shape of a horseshoe in two dimensions, while CA produced an arch. The arrangement of the sites along these bent structures, which indicates a replacement of species along an environmental gradient (see discussions in ter Braak 1987*b* and Legendre and Legendre 1998), is essentially the same as in the polynomial canonical ordinations presented below (Figs. 4b and 5). Van der Aart and Smeenk-Enserink (1975) had selected environmental variables to explain this gradient. They tested their hypotheses using CCoA; ter Braak (1986) did the same using CCA. We will now show that polynomial RDA and CCA provide better tests of these hypotheses than the linear forms. The results are better in two ways: the polynomial analyses provide (1) a higher proportion of variation of the species data explained by the model, which leads to more significant statistical tests, and (2) clearer identification of the variables explaining the gradient.

TABLE 1. Extended.

Sp. 9, *Pa. nigriceps*	Sp. 10, *Pa. pullata*	Sp. 11, *Tr. terricola*	Sp. 12, *Zo. Spinimana*	Water content	Reflection	*Calamagrostis* coverage	*Corynephorus* coverage
12	45	57	4	10.3	50	50	0
15	37	65	9	21.1	5	80	0
18	45	66	1	12.9	40	30	0
29	94	86	25	14.5	20	100	0
135	76	91	17	20.4	10	90	0
27	24	63	34	29.4	2	10	0
89	105	118	16	24.0	10	90	0
2	1	30	3	13.8	2	10	0
1	1	2	0	12.0	30	0	20
0	0	1	0	9.0	40	0	20
0	1	4	0	9.2	40	0	30
1	8	13	0	9.9	40	2	50
53	72	97	22	33.7	30	80	0
15	72	94	32	21.9	3	20	0
0	0	25	3	26.3	2	0	0
2	0	28	4	20.7	1	0	0
0	0	23	2	28.0	3	0	0
0	0	25	0	22.7	3	0	0
1	0	22	3	18.6	1	0	0
0	0	22	2	22.4	1	0	0
0	1	18	2	19.6	1	0	0
0	0	1	0	3.5	50	2	2
0	0	1	0	3.3	60	2	20
0	0	0	0	5.2	55	2	20
6	0	16	6	6.2	10	1	0
0	0	1	0	2.7	80	0	10
0	0	0	0	2.6	40	0	20
0	0	2	0	2.6	40	0	30

Example one: polynomial redundancy analysis (RDA)

The 12 species of spiders form the matrix of response variables **Y** (Table 1). In order to keep our first example small and manageable, the matrix of explanatory variables **X** only contains the first two environmental variables: water content of the soil (variable ''Water'') and reflection of soil surface (variable ''Reflection;'' high values of reflection indicate dry sites). The two environmental variables are highly negatively correlated in the data set: $r = -0.7482$. The species data were $\log_e(y + 1)$ transformed before analysis. Centering the explanatory variables on their respective means, before calculating the quadratic terms of the polynomial, reduced the collinearity between the linear and quadratic terms, as explained in step 1 of the polynomial regression algorithm.

The eigenvectors (species scores from Eq. B.4) were normalized to length 1 in order to represent the species and sites as a distance biplot. The site scores which

TABLE 2. Polynomial regression modeling of the spider species data (log-transformed variables) with respect to water content and reflection of the soil.

Species	Water	Reflection	$(\text{Water})^2$	Water × Reflection	$(\text{Reflection})^2$	R^2
Sp. 1	−0.28	0.82			−0.47	0.8267
Sp. 2	0.75	0.69		0.67		0.5524
Sp. 3	−0.84		0.54			0.8440
Sp. 4	0.47					0.2230
Sp. 5	−0.50		0.31	−0.29	0.30	0.8981
Sp. 6				0.44		0.1943
Sp. 7		−0.84		0.50	0.80	0.6147
Sp. 8		0.72			−0.72	0.4873
Sp. 9	0.45					0.2049
Sp. 10	0.77	0.65		0.45		0.3795
Sp. 11	0.81		−0.29			0.6720
Sp. 12	0.67					0.4526

Notes: The table gives only the standard partial regression coefficients (which are comparable) for the terms that were selected by backward elimination (rejection level: $\alpha = 0.05$). The intercepts are omitted. ''Water'' refers to the water content of the soil; ''reflection'' refers to the reflection of soil surface.

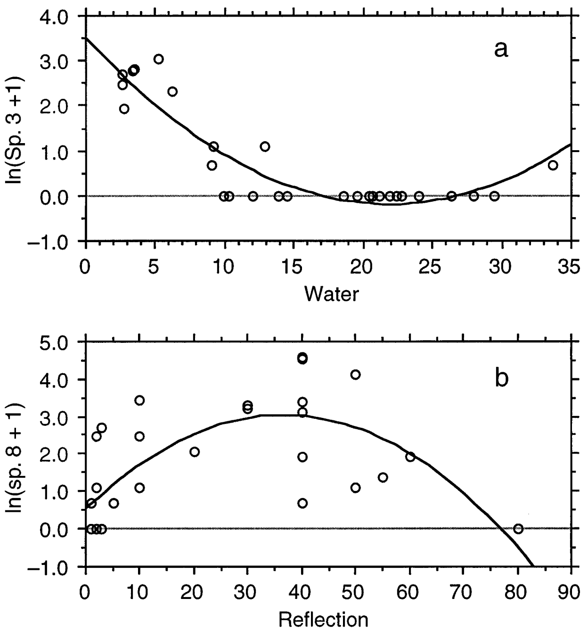

FIG. 3. Two examples of quadratic relationships of spider species from Table 1, after $\log_e(y + 1)$ transformation. For the upper panel the regression equation for uncentered variables ($R^2 = 0.8440$) is ln(Sp. 3 + 1) = 3.4984 − 0.3364 × (Water) + 0.0077(Water)^2, and the regression equation for centered variables ($R^2 = 0.8440$) is ln(Sp. 3 + 1) = −0.6198 − 0.1027(Water) + 0.0077(Water)^2. For the lower panel the regression equation for uncentered variables ($R^2 = 0.4873$) is ln(Sp. 8 + 1) = 0.5366 − 0.1368(Reflection) + 0.0019 × (Reflection)^2 and the regression equation for centered variables ($R^2 = 0.4873$) is ln(Sp. 8 + 1) = 0.9477 − 0.0475 × (Reflection) + 0.0019(Reflection)^2. Sp. 3 is shown as a function of water content whereas Sp. 8 is shown with respect to reflection of soil surface. The standard partial regression coefficients are given in Table 2. The R^2 coefficient is the same for noncentered and centered data although the equations differ.

are combinations of the environmental variables **X** were obtained from Eq. 6. Biplot scores for the two environmental variables, in polynomial form, were obtained from Eq. 12 using multiple linear correlations. In addition, simple linear correlations were computed for the individual terms of the quadratic polynomial. The correlations were scaled using Eq. B.7 for representation in the biplot. A linear RDA was also computed for comparison (Table 3).

Permutation tests were performed for linear and polynomial RDA to assess the significance of the two models. In both cases, the *P* value was 0.001 after 999 permutations; the two models were highly significant. The significance of the difference in explained variation between the two models was assessed using the test described in the section *Tests of significance in polynomial RDA and CCA*; the *P* value was 0.002 after 999 permutations. So the polynomial model seems more appropriate than the linear model to describe the relationships between **Y** and **X**. Detailed results of the polynomial RDA are presented in Table 4.

After hypothesis testing, one may be interested in looking at the species–environment relationships in some detail. Consider the first species (Sp. 1, *Al. accentuata*) of Table 1, for example. The polynomial regression algorithm provided the following quadratic equation to approximate the abundances (log-transformed) at the various sites *i*:

$$\hat{y}_i\ (\text{Sp. 1}) = 0.3585 - 0.0528x_{i1} + 0.0392x_{i2} + 0.0022x_{i1}^2 + 0.0006x_{i1}x_{i2} - 0.0009x_{i2}^2$$

where $\mathbf{x}_1$ is Water and $\mathbf{x}_2$ is Reflection. With only two explanatory variables, as in the present example, our polynomial regression algorithm makes no selection among the five terms of the quadratic polynomial equation. Because there are only two explanatory variables in the analysis, the same equation would have been obtained using the linear and quadratic variables as explanatory variables in a regular multiple regression. With more variables, our polynomial regression algorithm does not guarantee that the terms selected in the equation always represent the most optimal combination; this is the case for any step-by-step variable reduction procedure.

To appreciate the advantages of polynomial RDA, compare the biplots obtained from the linear and polynomial analyses (Fig. 4). Biplot 4a is from the linear RDA. Biplot 4b corresponds to polynomial RDA. The sites are positioned in terms of their responses to the explanatory variables in the biplot (Eqs. B.6 and 6) because the site scores are linear combinations of the environmental variables.

1) Polynomial RDA produced five canonical axes (Table 4) explaining 57.6% of the variation of **Y**. A large portion of the variance (53.7%) is accounted for by the first two canonical axes. This is considerably more than the 35.4% of the variation of **Y** accounted for by linear RDA on two canonical axes. The difference is due to the fact, shown in Table 2 (see also Fig. 4b), that most species (all except Sp. 4, 9, and 12) are

→

FIG. 4. RDA distance biplots of the spider species data of Table 1: results of (a) linear and (b) polynomial RDA. The numerical results of the polynomial RDA are in Table 4. The sites scores are linear combinations of the environmental variables. Dots are the sampling sites (with site numbers). Full lines without arrowheads represent the species. Full arrows represent the biplot scores of environmental variables for the individual terms of the polynomial; dashed arrows represent the biplot scores of environmental variables based upon the multiple correlations of (Water, [Water]^2) and (Reflection, [Reflection]^2) with the axes (Eqs. 12 and 13). The lengths of all species lines and environmental variable arrows have been multiplied by 10 for clarity; this does not change the interpretation of the biplots.

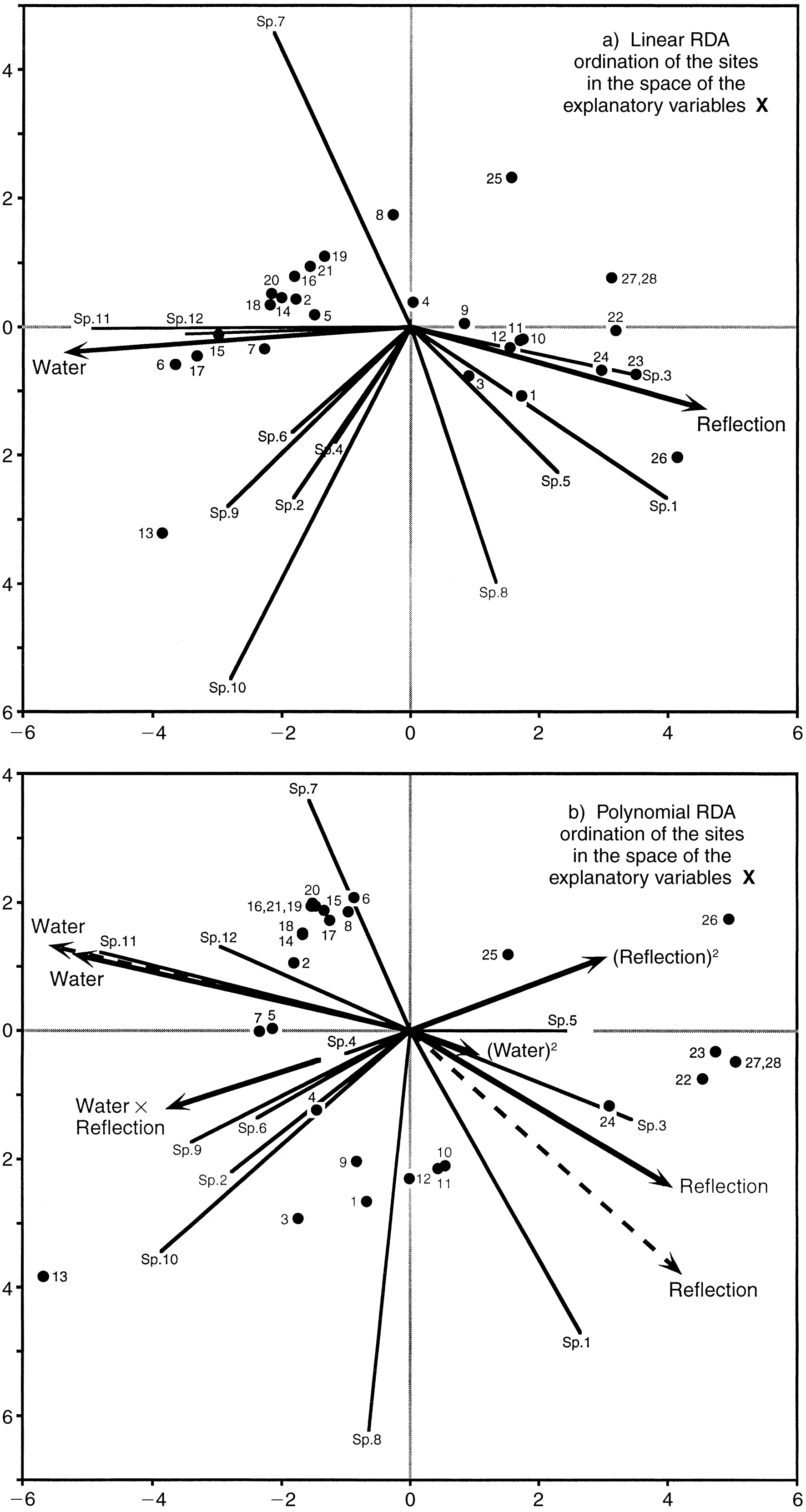
a) Linear RDA
ordination of the sites
in the space of the
explanatory variables X
Water
Reflection
Sp.1
Sp.2
Sp.3
Sp.4
Sp.5
Sp.6
Sp.7
Sp.8
Sp.9
Sp.10
Sp.11
Sp.12
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27,28
b) Polynomial RDA
ordination of the sites
in the space of the
explanatory variables X
Water
Water
(Reflection)2
(Water)2
Water ×
Reflection
Reflection
Reflection
16,21,19
−6
−4
−2
0
2
4
6

TABLE 3. Canonical eigenvalues and related data obtained using linear RDA for the spider species data.

	Canonical axes	
	I	II
Eigenvalues (with respect to total variance in **Y** = 20.36706)		
	6.04197	1.16368
Fraction of total variance in **Y**		
	29.66540	5.71352
Cumulative fraction of total variance in **Y** accounted for by axes I and II		
	29.66540	35.37892

related significantly to the quadratic terms of the polynomial of the explanatory variables: $(\text{Water})^2$, Water × Reflection, and $(\text{Reflection})^2$.

2) The positions of the species with respect to the environmental variables Water and Reflection are mostly the same in the linear and polynomial biplots, except for species 5. One diagram is simply rotated by ~20° with respect to the other. The angular order of the species around the biplot is almost identical to that in the PCA species diagram presented in Fig. 3 of van der Aart and Smeenk-Enserink (1975). So this is not where we should look for differences between the linear and polynomial solutions.

3) As mentioned above, the PCA solution published by van der Aart and Smeenk-Enserink (1975: Fig. 4) had the shape of a horseshoe; in PCA ordination, when there is replacement of the species along a single gradient, the ordination of the sites has a horseshoe shape in two dimensions. In the linear RDA biplot (Fig. 4a), the sites are shrunk into a crescent because the analysis is trying, with little success, to model their positions as linear responses to the two environmental variables; the linear analysis is not very successful at reconstructing the gradient. In the polynomial RDA biplot on the contrary (Fig. 4b), the sites are distributed in the same horseshoe fashion as in the PCA ordination. The species–environment correlation of polynomial RDA are 83% and 82%, respectively, for canonical axes 1 and 2. For linear RDA using Water and Reflection, these correlations were 80% and 43%, respectively, for axes 1 and 2. Thus polynomial RDA has produced an important gain in accuracy of the representation of the sites, compared to linear RDA. The good reconstruction of the sites in biplot 4b is due to the presence of the quadratic terms of the environmental variables; they are needed to correctly model the species (Table 2) and obtain a horseshoe-like distribution of the sites.

4) In RDA biplots, projecting a site at right angle on a species approximates the value of the site along that species axis. It is easy to check, in Table 1, that the sites found in quadrant III of the biplot (Fig. 4b) have the highest frequencies of occurrence of the species found in that quadrant (species 2, 4, 6, 8, 9 and 10). The reconstructed site scores in the linear biplot (Fig. 4a) do not position these sites correctly with respect to those species.

5) In RDA biplots, the angles between the species and the environmental variables reflect their correlations. Indeed, the variable Water × Reflection has strong positive correlations with species 2, 6, 9, 10, 11, and 12 and a strong negative correlation with species 5. $(\text{Reflection})^2$ is strongly positively correlated only to species 5; $(\text{Water})^2$ has strong positive correlations only with species 3 and 5.

Example two: polynomial canonical correspondence analysis (CCA)

For CCA, matrix **Y** contained the same 12 species of spiders. The data were not log transformed because CA and CCA are designed to analyze frequency data directly. **Y** was transformed into matrix $\bar{\mathbf{Q}}$ of contributions to chi-square. The matrix of explanatory variables **X** contained all four environmental variables of Table 1. CCA based on polynomial regression was computed for these data. The results of the analysis were compared with those of classical linear CCA (Table 5). For the biplot, only the positions of the first-degree terms, their squares and the simple products were computed. The correlations of more complex terms with the ordination vectors (Eq. C.10 or C.11) could easily be computed, but their interpretation would be difficult.

Permutation tests were performed for linear and polynomial CCA to assess the significance of the two models; the rows of matrix $\bar{\mathbf{Q}}$ were randomized with respect to matrix **X** of the explanatory variables. In both cases, the P value was 0.001 after 999 permutations; so, the two models were highly significant. The significance of the difference in variance accounted for by the two models was assessed using a permutation test. The P value was 0.001 after 999 permutations; this strongly suggests that the polynomial model is more appropriate than the linear in this example. Detailed results of the analysis are the following:

1) The analysis produced 12 canonical axes. The corresponding canonical eigenvectors accounted together for 80.3% of the variation of $\bar{\mathbf{Q}}$. The first six axes are shown in Appendix E; they account together for 78.2% of the variation of $\bar{\mathbf{Q}}$. This is noticeably larger than the 43.8% of the variation of $\bar{\mathbf{Q}}$ accounted for on four canonical axes by CCA based on weighted linear regression. The first two canonical axes explain 52.2% of the variation and the first three 64.9%. Therefore, two or three dimensions would form interesting ordination spaces for biplots since these axes account for a great deal of the variation of $\bar{\mathbf{Q}}$. The higher fraction of explained variation obtained by polynomial CCA is the result of (1) the higher number of constrained ordination axes and (2) the inclusion of sec-

TABLE 4. Results of polynomial RDA of the spider species data (selected output).

	Canonical axes				
	I	II	III	IV	V
Eigenvalues (with respect to total variance in **Y** = 20.36706)					
	7.51597	3.42170	0.33258	0.31649	0.14770
Fraction of total variance in **Y**					
	36.90258	16.80017	1.63295	1.55395	0.72518
Cumulative fraction of total variance in **Y** accounted for by axes I–V					
	36.90258	53.70275	55.33570	56.88965	57.61483
Species scores (normalized eigenvectors, matrix **U**)					
Al. accentuata	0.26607	−0.46920	−0.22110	0.13133	−0.00433
Al. cuneata	−0.27671	−0.21817	−0.16059	0.25158	0.41811
Al. fabrilis	0.34661	−0.13584	−0.02597	0.61271	−0.15820
Ar. lutetiana	−0.09945	−0.03477	0.36491	0.06704	−0.28261
Ar. perita	0.24240	0.00115	0.44307	0.23270	0.67117
Au. albimana	−0.23744	−0.13512	−0.12822	0.27683	0.09486
Pa. lugubris	−0.15772	0.35779	−0.56306	0.06533	0.31349
Pa. monticola	−0.06342	−0.62076	−0.29975	−0.20388	−0.09266
Pa. nigriceps	−0.33866	−0.17318	0.03770	0.31161	−0.02935
Pa. pullata	−0.38561	−0.34346	0.39806	−0.22958	0.16962
Tr. terricola	−0.47841	0.12382	−0.02433	0.02489	0.00236
Zo. spinimana	−0.29378	0.13298	0.10671	0.46097	−0.35256
Site scores from Eq. B.5, vector **cord**					
Site 1	−2.35239	−3.76678	−0.18112	−0.30880	0.26217
Site 2	−3.44760	0.30630	1.15077	0.61495	0.15753
Site 3	−2.54619	−3.44231	−0.15865	0.61822	0.71142
Site 4	−4.47463	−1.51194	1.12045	1.10953	−0.03697
Site 5	−4.99663	−1.07440	1.52590	1.65727	0.36330
Site 6	−3.90795	−0.89575	1.75764	0.98327	−1.09182
Site 7	−5.13140	−2.64051	1.45755	1.01480	−0.53534
Site 8	−1.72782	2.21318	−2.21051	0.43691	1.48668
Site 9	1.60351	−0.64764	−0.30795	−2.03296	−0.43313
Site 10	2.92617	−0.50835	−0.61202	−1.72588	−0.91893
Site 11	2.28290	−2.47035	−1.30393	−1.35677	−0.64618
Site 12	0.23639	−3.19603	−0.91572	−1.83007	0.51147
Site 13	−4.83237	−2.83936	0.48656	1.94347	0.35610
Site 14	−4.09718	−0.49444	1.02181	0.20035	−0.49018
Site 15	0.57466	3.38082	−0.35775	−0.86334	−0.38539
Site 16	−0.29880	2.93130	−0.51779	−0.06665	0.08358
Site 17	0.83111	3.02950	0.09058	−1.05330	−0.54977
Site 18	0.81444	2.99010	−0.53020	−1.33807	0.34486
Site 19	0.10914	3.31586	−0.78850	−0.43557	0.07786
Site 20	0.30451	3.33570	−0.95217	−0.67733	0.39239
Site 21	0.16624	2.80162	−0.92359	−1.07187	0.33662
Site 22	4.37169	0.14199	0.50280	0.46620	0.00844
Site 23	4.45336	−0.36262	−0.33924	0.54394	0.59587
Site 24	4.91037	−0.34115	0.41907	0.61543	0.06733
Site 25	0.95260	−0.61796	−1.07489	1.59368	−1.24402
Site 26	4.27300	1.27126	1.59992	0.36615	1.14963
Site 27	4.95996	−0.26499	0.35801	0.49178	0.09376
Site 28	4.04290	−0.64305	−0.31703	0.10466	−0.66730
Biplot scores of environmental variables (from Eq. 12)					
Water	−0.55831	0.13356	0.08258	0.09274	−0.02380
Reflection	0.42253	−0.37818	0.04846	0.01118	0.06100
Biplot scores of environmental variables (from Eq. B.7)					
Water	−0.52272	0.12139	0.04417	0.01535	−0.01643
Reflection	0.40913	−0.24304	0.02803	0.00474	0.03246
(Water)^2	0.11014	−0.03559	0.07593	0.09274	−0.01962
Water × Reflection	−0.37820	−0.12150	−0.08382	0.01591	−0.02378
(Reflection)^2	0.30721	0.11586	0.04838	0.01109	0.06100

Notes: Matrix **Y**: hunting spider species 1–12. Matrix **X**: water content, reflection of soil surface. Either set of biplot scores can be used to represent the environmental variables in biplots. Users of the program may also request the matrix of regression coefficients **B** of the multiple linear regressions of **Y** on **X** (if classical linear RDA or CCA is computed) or the polynomial coefficients for each response variable **y** of **Y** (if polynomial RDA or CCA is used). The program may also carry out permutation tests of the significance of the linear and polynomial models, as well as the significance of the difference in variance accounted for between the two models.

ond-order terms which allow full recovery of the CA arch.

2) The linear CCA solution produced quite a bit of distortion to the arch representing the gradient in the CA biplot, because it imposed the constraint that the ordination axes be linearly related to the environmental variables. The linear CCA ordination is not shown here; the positions of the points are similar to the CCA results presented in Fig. 1 of ter Braak (1986) with some differences due to the fact that only two of the environmental variables (water content and reflection of soil suface) were the same in the two analyses. The polynomial CCA solution (Fig. 5) is more successful at recovering the arch because it incorporates quadratic environmental terms in the explanatory equations of the species. As a result, the ordination of the sites in Fig. 5 is very similar to that of the CA biplot. The biplot is dominated by the opposition between two pairs of environmental variables in linear and quadratic forms: on the one hand, water content and cover by the grass *Calamagrostis epigejos* indicate wet sites, which are found in quadrant II of Fig. 5. On the other hand, reflection of soil surface and cover by the grass *Corynephorus canescens* indicate dry sites; abundance of *Corynephorus* was highly correlated with the percentage of bare sand in van der Aart and Smeek-Enserink (1975). In the linear CCA biplot (*not shown*), which does not display the arch properly, *Calamagrostis* is not associated with water content, and *Corynephorus* is not associated with reflection of the soil surface.

3) The sites form three main groups, more densely clustered than in the RDA ordination (Fig. 4): the driest sites 22 to 28, found in quadrant I of Fig. 5, are associated with high frequencies of species 3 and 5; the more humid sites 2, 4 to 8, and 13 to 21 (in the insert of Fig. 5) are associated with high frequencies of species 4, 6, 7, and 9 to 12; sites 1, 3, and 9 to 12, with intermediate humidity, are associated with high values of species 8.

4) Projecting the species at right angles on the water content variable, for example, provides an ordination of the species of spiders along this variable. Sp. 5 has the lowest weighted average with respect to water content, followed by Sp. 3, Sp. 1, and Sp. 8; all the other species (except Sp. 2), found in quadrant II of the biplot, occupy approximately the same position on the positive side of this variable. When projecting the spider species onto the *Corynephorus* percent cover, they clearly fall into two groups; species 1, 3, 5, and 8, mentioned in the previous sentence, occupy nearly the same position along this variable.

5) The environmental variables were centered before the other terms of the polynomial expression were computed. In CCA, the centering involves the row weights p_{i+} of the species data table. This means, for instance, that high values of the Water × Reflection variable would correspond to sites having high (or low) values for both variables; no such site is found in the data set, with the consequence that none occupies quadrant III where this variable is pointing, except site 3 which lies near the origin. Sites 22 to 28 have, however, high negative values for this product variable, due to the very low water content combined with high values of reflection of the soil; so they are found in quadrant I, which is opposite to the arrow representing this product variable. Sites 22 to 28 also have high values of $(\text{Water})^2$ (because they have the most extreme values of the centered variable Water, on the negative side) and $(\text{Reflection})^2$ (because they have among the highest values of centered Reflection, on the positive side). *Calamagrostis* and *Corynephorus* are both absent from sites 15 to 21, found high in the insert of Fig. 5; as a consequence, these sites have the highest negative values on both of these centered variables, which gives them the highest positive values for the product variable *Calamagrostis* × *Corynephorus*. There are no sites where both of these plant species are found together in any abundance. The role of the other product variables in the analysis can be interpreted in the same way.

In this biplot, the individual terms of the polynomial as well as the combined terms have been drawn in order to show how polynomial CCA allowed the full representation of the arch. In an actual application of the method, a simpler diagram showing only the arrows for the multiple correlation biplot scores (Water and $[\text{Water}]^2$ combined, etc.) would be sufficient to describe the main environmental axes of variation of the data.

The equations generated by the polynomial regression algorithm to approximate the $\bar{\mathbf{q}}$ values of the first species (Sp. 1: *Al. accentuata*) are the following:

$$\begin{aligned} x_{i14} &= -0.4167 - 0.0464x_{i1} - 0.0708x_{i4} \\ &\quad + 0.000241x_{i1}^2 - 0.007547x_{i1}x_{i4} \\ &\quad - 0.000091x_{i4}^2 \\ x_{i23} &= 0.1440 - 0.0030x_{i2} + 0.0075x_{i3} \\ &\quad - 0.000034x_{i2}^2 - 0.000071x_{i2}x_{i3} \\ &\quad - 0.000390x_{i3}^2 \\ \hat{q}_i\ (\text{Sp. } 1) &= x_{i,1423} \\ &= -0.0101 + 1.0157x_{i14} + 0.9915x_{i,23} \\ &\quad + 0.508529x_{i14}x_{i23} \end{aligned}$$

where $\mathbf{x}_1$ is water content, $\mathbf{x}_2$ is reflection of the soil surface, $\mathbf{x}_3$ is percent cover by *Calamagrostis*, and $\mathbf{x}_4$ is percent cover by *Corynephorus*. The three equations above illustrate the approximation process for the first species: in the first iteration, the explanatory variables $\mathbf{x}_1$ and $\mathbf{x}_4$ were combined to form a new variable $\mathbf{x}_{14}$ (first equation), $\mathbf{x}_2$ and $\mathbf{x}_3$ in the second iteration to form $\mathbf{x}_{23}$ (second equation), the new combined variables $\mathbf{x}_{14}$

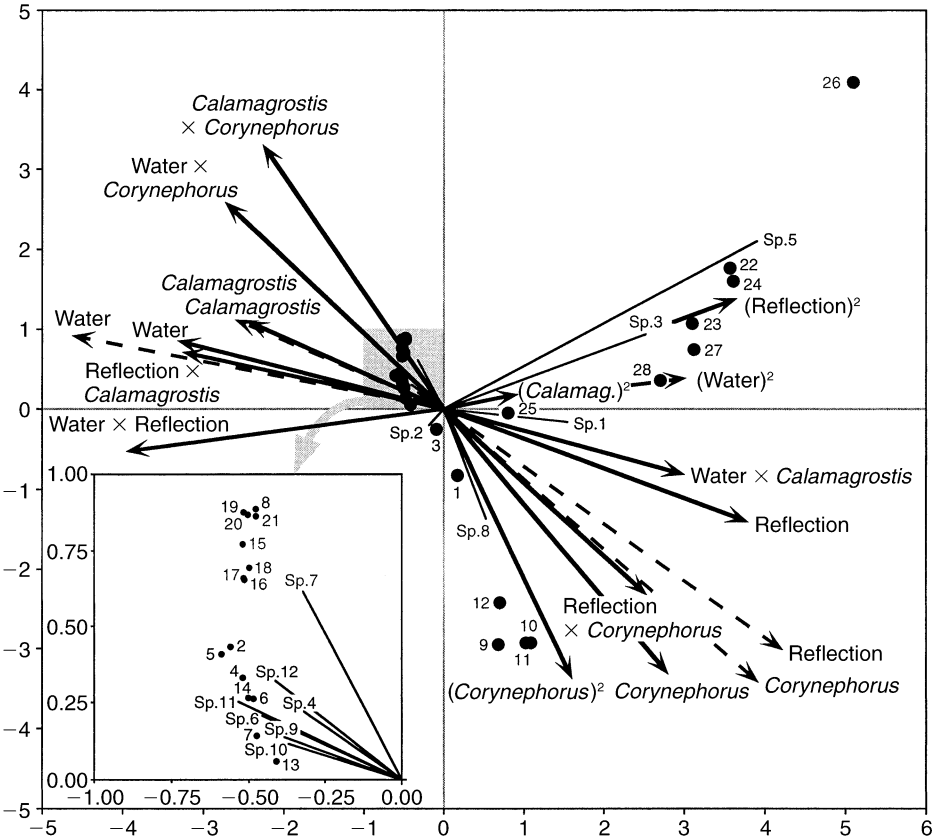

FIG. 5. Polynomial CCA biplot (scaling type 2) for the spider species data presented in Table 1; the numerical results of the analysis are in Appendix E. Dots are the sampling sites (sites scores are from matrix $\hat{\mathbf{V}}$, Eq. C.7 [see Appendix C]); numbers are the site numbers. Full lines without arrowheads represent the species (species scores are from matrix $\hat{\mathbf{F}}$, Eq. C.9). Full arrows represent the biplot scores of environmental variables for the individual terms of the polynomial (Eq. C.12); dashed arrows represent the biplot scores of environmental variables based upon the multiple correlations of the linear and quadratic terms with the axes (Eqs. 12 and 13). The lengths of the environmental variable arrows have been multiplied by five for clarity; this does not change the interpretation of the diagram. The insert shows details of the ordination of the species and sites in quadrant II.

and $\mathbf{x}_{23}$ were joined in the third iteration to form $\mathbf{x}_{1423}$ (third equation), which is equal to the estimated value of the response variable $\bar{\mathbf{q}}$. The development of this regression procedure is depicted in Fig. 2.

The polynomial model does not necessarily provide such good results for all data sets; there are indeed cases where the response variables in $\mathbf{Y}$ are only linearly related to the explanatory variables. Using the principle of parsimony of the 14th century logician and philosopher William Ockham, ''pluralites non est ponenda sine necessitate,'' the linear model must be seen as the best representation of the data, in such cases, because it contains fewer parameters. Our test of significance of the difference between the two models points users towards the most appropriate one.

TABLE 5. Canonical eigenvalues obtained using linear CCA for the data in Table 1.

	Canonical axes			
	I	II	III	IV
Eigenvalues (with respect to total variance in $\bar{\mathbf{Q}}$ = 1.92296)				
	0.54518	0.17247	0.09789	0.02682
Fraction of total variance in $\bar{\mathbf{Q}}$				
	28.35114	8.96922	5.09045	1.39477
Cumulative fraction of total variance in $\bar{\mathbf{Q}}$ accounted for by axes I–IV				
	28.35114	37.32036	42.41081	43.80558

DISCUSSION

Researchers often want to test hypotheses relating response (e.g., species data) to explanatory (e.g., environmental) variables; canonical analysis is appropriate in such studies. In many instances, the hypotheses do not specify that the relationships between the two data sets are linear; they are not, in most cases, when analyzing species composition data. We have described how redundancy analysis (RDA) and canonical correspondence analysis (CCA) can be modified to express polynomial relationships between the response ($\mathbf{Y}$) and explanatory variables (**X**), instead of linear relationships as in classical RDA and CCA.

An empirical polynomial regression algorithm was developed to do so. Consider a canonical analysis problem with a fairly small number of environmental variables, e.g., 10. The number of combination terms containing these variables in the first and second degree is very large. It may often be greater than the number of observations; this, in turn, would jeopardize the inversion $[\mathbf{X}'\mathbf{X}]^{-1}$ required to estimate the regression coefficients. Methods of selection of the most important terms of the polynomial equation are required to avoid overfitting. The problem can be approached from two angles. (1) The first angle is to reduce the number of variables in the model. Users of the method who are considering many explanatory variables may compute polynomial RDA or CCA with different combinations of explanatory variables to discover the combination providing the most significant polynomial model. In fact, some data may be better explained by including linear and quadratic contributions of some variables and only linear contributions of the others. Furthermore, when working with m explanatory variables, at any iteration number k ($1 \leq k < m - 1$), one could check the level of significance of the intermediate linear regression of $\mathbf{y}$ on the reduced matrix $\mathbf{X}$ comprised of $m - k$ columns. Such a strategy would allow users to select an intermediate regression model which would be neither a classical linear model nor a complete quadratic polynomial. Our polynomial regression algorithm could easily be adapted to accommodate these modifications. (2) When the variables to include in the model have been selected, the second angle is to reduce the number of terms (combinations of the original variables) in the model. This could be done in a variety of ways, all of which would be heuristic. Our method contains a heuristic selection strategy meant to optimize the least-squares loss function, at each step and also in general. Actually, the algorithm performs a number of linear multiple regressions one after the other. The loss function minimized by the method is the same as in classical multiple regression. In our algorithm, each variable is limited to a power of two in any term of the polynomial equation. Like any heuristic procedure, this one may find a local minimum instead of the global one; its main advantage is that it runs in polynomial time with respect to the size of the data matrices, whereas a procedure that would try in turn all possible subsets of the full polynomial model would be running in exponential time and would thus be inapplicable to real data sets. Our recommendation to users is to use polynomial RDA or CCA on data sets containing more than ($3m - 1$) observations.

Polynomial regression does not guarantee to always produce a model with greater significance than the linear model. If both the linear and polynomial models prove to be significant, a permutation test may be used to assess the difference in variance accounted for by the two models and determine which is the most appropriate one to describe the data. In the real-data examples reported in this paper, the polynomial models of the explanatory variables fitted to the data were demonstrably better than the linear models. From the ecological point of view, they fitted the horseshoe or arch representing the gradient present in the data much more efficiently than the linear forms of analysis. From the statistical point of view, they accounted for greater percentages of the total variance of the response variables than classical RDA and CCA based upon linear regression, and explained a significant part of the variation which had remained unexplained by the linear models. On the other hand, simulations have shown that if the response variables are linearly related to the explanatory variables, the test of significance of the difference in explained variation will point to the linear canonical model as being the most appropriate; if the response-to-explanatory relationships are polynomial, the test will point to the polynomial model as the most appropriate one.

Acknowledgments

We benefited from very interesting comments by R. H. Økland and T. Philippi, as well as another anonymous reviewer, to whom we are grateful. This research was supported by NSERC grant No. OGP7738 to P. Legendre. V. Makarenkov was suported in part by a postdoctoral fellowship from the Group for Interuniversity Research in Limnology and Aquatic Environment (GRIL).

Literature Cited

Donovan, C. 1998. Optimal transformations in multivariate analysis of trawl data. Thesis. University of Auckland, Auckland, New Zealand.

Durand, J. F. 1993. Generalized principal component analysis with respect to instrumental variables via univariate spline transformations. Computational Statistics and Data Analysis **16**:423–440.

Gabriel, K. R. 1982. Biplot. Pages 263–271 *in* S. Kotz and N. L. Johnson, editors. Encyclopedia of statistical sciences. Volume 1. Wiley, New York, New York, USA.

Legendre, P., and L. Legendre. 1998. Numerical ecology. Second English edition. Elsevier, Amsterdam, The Netherlands.

Økland, R. H. 1999. On the variation explained by ordination and constrained ordination axes. Journal of Vegetation Science **10**:131–136.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhyaá **A 26**: 329–358.

Rao, C. R. 1973. Linear statistical inference and its applications. Second edition. Wiley, New York, New York, USA.

ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology **67**:1176–1179.

ter Braak, C. J. F. 1987*a*. The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio **69**:69–77.

ter Braak, C. J. F. 1987*b*. Ordination. Pages 91–173 *in* R. H. G. Jongman, C. J. F. ter Braak, and O. F. R. van Tongeren, editors. Data analysis in community and landscape ecology. (Reissued in 1995 by Cambridge University Press, Cambridge, UK.) Centre for Agricultural Publishing and Documentation, Wageningen, The Netherlands.

ter Braak, C. J. F. 1988*a.* CANOCO: an extension of DECORANA to analyze species-environment relationships. Vegetatio **75**:159–160.

ter Braak, C. J. F. 1988*b.* CANOCO: a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis. Version 2.1. Agricultural Mathematics Group, Ministry of Agriculture and Fisheries, Wageningen, The Netherlands.

ter Braak, C. J. F. 1990. Update notes: CANOCO version 3.10. Agricultural Mathematics Group, Wageningen, The Netherlands.

ter Braak, C. J. F. 1994. Canonical community ordination. Part I: basic theory and linear methods. Écoscience **1**:127–140.

ter Braak, C. J. F., and P. Smilauer. 1998. CANOCO reference manual and user's guide to CANOCO for Windows: software for canonical community ordination. Version 4. Microcomputer Power, Ithaca, New York, USA.

van der Aart, P. J. M., and N. Smeenk-Enserink. 1975. Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. Netherlands Journal of Zoology **25**:1–45.

van der Burg, E., and J. de Leeuw. 1983. Non-linear canonical correlation. British Journal of Mathematical and Statistical Psychology **36**:54–80.

APPENDIX A

A more detailed consideration of the issues of the number of degrees of freedom and the number of independent parameters involved in the computation of the polynomial regression procedure is available online in ESA's Electronic Data Archive: *Ecological Archives* E083-018-A1.

APPENDIX B

A description of the direct computational approach to redundancy analysis (RDA) is available online in ESA's Electronic Data Archive: *Ecological Archives* E083-018-A2.

APPENDIX C

A description of the direct computational approach to canonical correspondence analysis (CCA) is available online in ESA's Electronic Data Archive: *Ecological Archives* E083-018-A3.

APPENDIX D

A description of the permutational methods used in polynomial RDA and CCA to test the significance of the relationships between the response and explanatory data matrices, and to assess the difference in variance accounted for between the polynomial model and the linear model nested into it, is available online in ESA's Electronic Data Archive: *Ecological Archives* E083-018-A4. The appendix also reports the results of simulations showing that the tests have correct type I error.

APPENDIX E

A table of results of polynomial CCA of the spider data (selected output) is available online in ESA's Electronic Data Archive: *Ecological Archives* E083-018-A5.

SUPPLEMENT

Software to compute nonlinear canonical analysis (program POLYNOMIAL RDACCA: source code, compiled versions for Macintosh and Windows, program documentation, and example data files) is available online in ESA's Electronic Data Archive: *Ecological Archives* E083-018-S1. Also available on the WWWeb site <http://www.fas.umontreal.ca/biol/legendre/>.

Appendices to:

Makarenkov, V. and P. Legendre. 2002. Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. *Ecology* 83(4): 1146-1161.

APPENDIX A

ECOLOGICAL ARCHIVES E083-018-A1

NUMBER OF INDEPENDENT PARAMETERS ESTIMATED BY THE POLYNOMIAL REGRESSION PROCEDURE

This Appendix will consider in more detail the issues of the number of degrees of freedom and the number of independent parameters involved in the computation. Since the new polynomial algorithm goes through a number of complex steps, this issue may be a source of confusion for non-statistically-oriented readers who may be afraid of overfitting the data. The large number of terms which can be obtained in the complete polynomial form does not reflect the number of *independent* parameters estimated by the polynomial algorithm. Consider the four (*m*) environmental variables and the first response variable (species 1) in Table 3 of the main paper, used to illustrate polynomial CCA in the second numerical example of the paper; *n* = 28.

• During the first iteration, in the first step of multiple linear regression (Eq. 1 of the main paper), *m* = 4 parameters were estimated (plus the intercept, which will not be mentioned again), one for each of the variables $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$.

• During step 3 of the algorithm, 5 parameters were estimated using Eq. 3; they pertain to $\mathbf{x}_1$, $\mathbf{x}_4$, $\mathbf{x}_1^2$, $\mathbf{x}_1\mathbf{x}_4$, and $\mathbf{x}_4^2$. Two of these parameters refer to variables $\mathbf{x}_1$ and $\mathbf{x}_4$. Since we already had initial estimates for these 2 parameters (from Eq. 1, previous paragraph), they were combined with the new estimates of these parameters obtained from Eq. 3. This provided the vector of fitted values called $\mathbf{x}_{14}$ (Eq. 4) which was calculated using a total of 5 parameters. Considering the parameters estimated above for $\mathbf{x}_2$ and $\mathbf{x}_3$, our total up to now is 7 parameters. The 5 parameter estimated during step 3 are dependent upon the estimates initially obtained for parameters $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$. Non-independent parameter estimates are discussed further below.

• During the second iteration, 5 parameters again were estimated to provide the vector of fitted values denoted $\mathbf{x}_{23}$; they belong to $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_2^2$, $\mathbf{x}_3^2$, and $\mathbf{x}_2\mathbf{x}_3$. Two of these pertain to $\mathbf{x}_2$ and $\mathbf{x}_3$ for which we already had initial estimates, with which they were combined using Eq. 4. Our total up to now is 10 parameters, but again, the group of 5 parameter estimates are dependent upon the parameters previously estimated for $\mathbf{x}_1$, $\mathbf{x}_4$, and $\mathbf{x}_{14}$.

• During the third and final iteration of the algorithm, 3 parameters were estimated during step 3 ($\mathbf{x}_{14}$ $\mathbf{x}_{23}$ and $\mathbf{x}_{14}\mathbf{x}_{23}$), but only one of them, for the cross-product $\mathbf{x}_{14}\mathbf{x}_{23}$, was a new parameter. The parameters involved in the construction of $\mathbf{x}_{14}$ and $\mathbf{x}_{23}$, in the previous 2 paragraphs, were not re-estimated during the last computation step.

The total number of parameters is 11 in this example. They include the 4 independent parameters initially computed for the linear terms, plus all the other parameters estimated during the following steps, that depended partly upon them. So, the number of degrees of freedom used by the equation to estimate the fitted values $\hat{q}_i$ (Sp. 1) is a fractional number between 4 and 11 in this example.

In our polynomial regression procedure, for a matrix of environmental variables with *m* columns, the actual number of estimated parameters is 3*m*–1; they are not all estimated independently of one another, however. Independent estimates are obtained when parameters are estimated simultaneously, as part of the same estimation process (e.g., when they pertain to the same regression equation). Each parameter estimate is a partial estimate which takes into account the values of all the other parameters present in the equation. The number of degrees of freedom used by the estimation process is given by the number of *independent parameter estimates.* This number can be fractional. When parameters are estimated from residuals of a model, the new estimates are not independent of the model parmeters used to compute these residuals. (Fractional numbers of degrees of freedom are found in other methods of data analysis used by ecologists, for instance in tests of significance for spatially autocorrelated data.) In our algorithm for polynomial regression, estimating some parameters in a non-independent way reduces the fit of the model to the data by some small amount, but it increases the power of the tests of significance. Some authors might prefer to use the opposite strategy, sacrificing power to precision; the problem is, however, that the polynomial regression procedure would be computationally much slower on presently available hardware, impairing the permutation tests described in the section “Tests of significance in polynomial RDA and CCA” even for fairly small data sets.

The actual number of degrees of freedom left for statistical testing is (*n*–1) minus the number of independent parameters. The number of independent parameters is a number larger than *m* and smaller than or equal to (3*m*–1); *m* is the number of independent parameters estimated during the first iteration of the algorithm; (3*m*–1) is the total number of parameters estimated during all cycles of the algorithm, including the independent and dependent parameters.

The maximum number of non-zero canonical eigenvalues and corresponding canonical axes that can be obtained in polynomial RDA and CCA is min[*p*,(*n*–1)]. Our recommendation to preserve good power in statistical tests is to only use polynomial RDA or CCA on data sets containing more than (3*m*–1) observations.

APPENDIX B

ECOLOGICAL ARCHIVES E083-018-A2

OUTLINE OF CLASSICAL REDUNDANCY ANALYSIS (RDA)

The direct computational approach to redundancy analysis proceeds as follows (Legendre and Legendre 1998). The mathematics behind RDA is summarized here to make it easier for readers to understand the modifications required by polynomial RDA. Matrices **Y** and **X** have been described in the section "Redundancy Analysis and its Polynomial Generalization".

1) Data preparation and multiple regression. For convenience, the variables in **Y** and **X** are centered on their respective means. The first step consists of carrying out multiple linear regressions for each variable in **Y** on all variables in **X** and computing the fitted values. The matrix of fitted values $\hat{\mathbf{Y}}$ used in the following steps is obtained from the equation:

$$\hat{\mathbf{Y}} = \mathbf{XB} = \mathbf{X}[\mathbf{X'X}]^{-1}\mathbf{X'Y} \qquad \text{(B.1)}$$

where **B** is the matrix of regression coefficients of the response variables **Y** on the regressors **X**. An extra column with 1's should be added to matrix **X**, before the multiple regression, to allow estimation of the intercepts. (In linear RDA, centering the variables in **Y** and **X** eliminates the intercepts; this is not necessarily the case in polynomial RDA. Centering **X** offers the additional advantage of reducing the collinearity between the linear and quadratic terms of the polynomial, as explained in step 1 of the polynomial regression algorithm.)

2) The covariance matrix **S** of the matrix of fitted values $\hat{\mathbf{Y}}$ is computed as follows:

$$\mathbf{S} = [1/(n-1)]\ \hat{\mathbf{Y}}'\ \hat{\mathbf{Y}} \qquad \text{(B.2)}$$

or, incorporating the development from Eq. B.1:

$$\mathbf{S} = [1/(n-1)]\ \mathbf{Y'X}[\mathbf{X'X}]^{-1}\mathbf{X'Y} \qquad \text{(B.3)}$$

3) Principal components of the table of fitted values $\hat{\mathbf{Y}}$ are computed to reduce the dimensionality of the solution. This corresponds to solving the eigenvalue problem:

$$(\mathbf{S} - \lambda_k\mathbf{I})\mathbf{u}_k = \mathbf{0} \qquad \text{(B.4)}$$

where λ_k denotes the *k*-th eigenvalue and $\mathbf{u}_k$ the associated eigenvector. The matrix containing the normalized canonical eigenvectors is called **U**. The eigenvectors give the contributions of the descriptors **Y** to the canonical axes. In linear RDA, matrix **U** is of size (p×$min[p, m, n - 1]$) because the number of canonical eigenvectors cannot exceed the minimum of *p, m* and ($n - 1$).

4) The canonical ordination of the objects (rows of **Y**) in the space of the response variables **Y** is obtained directly from the centered matrix **Y**, using the standard equation for principal components and eigenvectors $\mathbf{u}_k$ of Eq. B.4:

$$\mathbf{cord}_{(\text{space of response variable }\mathbf{Y})k} = \mathbf{Yu}_k \qquad \text{(B.5)}$$

The ordination vectors defined in this equations are called the vectors of "site scores"; Palmer (1993) calls them the "minimally constrained scores". These vectors have variances that are close but not equal to the corresponding eigenvalues.

Likewise, the canonical ordination of objects in the space of the explanatory variables **X** is obtained from the following formula:

$$\mathbf{cord}_{(\text{space of explanatory variables } \mathbf{X})k} = \hat{\mathbf{Y}}\,\mathbf{u}_k = \mathbf{XBu}_k \qquad \text{(B.6)}$$

In this case the ordination vectors are constrained linear combinations of the explanatory variables **X**. This is the reason why these constrained ordination vectors are also called "fitted site scores"; Palmer (1993) calls them the "maximally constrained scores". They have variances equal to the corresponding eigenvalues.

The "site scores" of Eq. B.5 are obtained by projecting the original data of matrix **Y** onto axis *k*; they approximate the observed data containing residuals ($\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{Y}_{\mathbf{res}}$). On the other hand, the "fitted site scores" of Eq. B.6 correspond to projecting the fitted values of matrix $\hat{\mathbf{Y}}$ onto axis *k*; they approximate the fitted data. Both sets can be used in biplots, as in Fig. 5 of the main paper.

5) The other important information needed for interpreting the relationships between the variables in **X** and **Y** is the contribution of the explanatory variables to the canonical axes. To assess this contribution, correlations are computed between the variables in **X**, on the one hand, and the canonical ordination axes in either space **Y** (Eq. B.5) or space **X** (Eq. B.6) on the other. The correlations between the variables in **X** and the canonical ordination axes in space **X** can be used to represent the explanatory variables in biplots.

6) In RDA, biplot diagrams can be drawn that contain three sets of data points: the site scores (from Eq. B.5 or B.6), the response variables from **Y**, and the explanatory variables from **X**. Each pair of sets of points forms a biplot. Biplots primarily serve to interpret the relationships between sites in terms of the **Y** and/or **X** variables. If there are too many sites or too many variables in **X** or **Y**, separate diagrams can be drawn and presented side by side. Two main types of scaling can be used in RDA biplots; for details on their properties and interpretation see ter Braak (1994) or Legendre and Legendre (1998). The biplot produced by type 1 scaling, called *distance biplot*, preserves the distances among sites. The biplot produced by type 2 scaling, called *correlation biplot*, focuses on the correlations among the response variables.

In RDA scaling of type 1, used in the "Numerical examples" section, the eigenvectors in matrix **U**, representing the response variable scores, are scaled to lengths 1. The fitted site scores from Eq. B.6 have variances equal to λ_k whereas the site scores from Eq. B.5 have variances which are usually slightly larger than λ_k. Each explanatory variable **x** can be represented in the biplot by means of the correlations of **x** with the fitted site scores. These correlations have to be multiplied by a coefficient c_k:

$$c_k = \sqrt{\lambda_k / \textit{Total variance in } \mathbf{Y}} \qquad \text{(B.7)}$$

where λ_k is the eigenvalue corresponding to axis *k*; this correction accounts for the fact that, in this scaling, the variances of the site scores differ among axes.

Literature Cited

Legendre, P., and L. Legendre. 1998. Numerical ecology. Second English edition. Elsevier, Amsterdam, The Netherlands.

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. 1994. Canonical community ordination. Part I: Basic theory and linear methods. Écoscience **1**: 127-140.

APPENDIX C

ECOLOGICAL ARCHIVES E083-018-A3

OUTLINE OF CLASSICAL CANONICAL CORRESPONDENCE ANALYSIS (CCA)

The direct computational approach to canonical correspondence analysis proceeds as follows (Legendre and Legendre 1998). The mathematics behind CCA is summarized here to make it easier for readers to understand the modifications required by polynomial CCA. Matrices **Y** and **X** have been described in the section “Canonical Correspondence Analysis and its Polynomial Generalization”.

1) Data preparation and multiple regression. Absolute frequencies (i.e., the individual species abundances in matrix **Y**) are represented by y_{ij} whereas relative frequencies (also called probabilities or proportions) by p_{ij}; p_{ij} is the frequency y_{ij} in cell *ij* divided by the sum y_{++} of the y_{ij}’s over the whole frequency table. Row weight p_{i+} is equal to y_{i+}/y_{++} where y_{i+} is the sum of values in row *i*. Likewise, column weight p_{+j} is equal to y_{+j}/y_{++} where y_{+j} is the sum of values in column *j*. CCA is computed from a matrix denoted $\bar{\mathbf{Q}}$ (n×p):

$$\bar{\mathbf{Q}} = [\bar{q}_{ij}] = \left[\frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}\right] \tag{C.1}$$

The values $\bar{q}_{ij}$ only differ by a constant from the contributions to chi-square (χ_{ij}) computed in cell *ij* of a contingency table during two-way contingency table analysis: $\bar{q}_{ij} = \chi_{ij}/\sqrt{y_{++}}$. This causes all the eigenvalues of $\bar{\mathbf{Q}}$ to be smaller than or equal to 1, as shown by Legendre and Legendre (1998, section 9.4). Values $\bar{q}_{ij}$ may be calculated directly from the y_{ij}’s:

$$\bar{q}_{ij} = \frac{y_{ij}y_{++} - y_{i+}y_{+j}}{y_{++}\sqrt{y_{i+}y_{+j}}} \tag{C.2}$$

The variables in **X** are centered on their respective means, as in step 1 of RDA and for the same reason, except that the means used to center **X** are computed here as the sums of the columns of $\mathbf{D}(p_{i+})\mathbf{X}$, where $\mathbf{D}(p_{i+})$ is a diagonal matrix containing the row weights p_{i+}.

As in RDA, multiple linear regression of $\bar{\mathbf{Q}}$ on the matrix of explanatory variables **X** is computed. Matrix **X** is weighted during this regression; the weights for the explanatory variables are given by a diagonal matrix, $\mathbf{D}(p_{i+})^{1/2}$, of the square roots of the rows weights of **Y**. The weighted matrix of explanatory variables, $\mathbf{X_w}$, is thus:

$$\mathbf{X_w} = \mathbf{D}(p_{i+})^{1/2}\mathbf{X} \tag{C.3}$$

The equation for the matrix of fitted values $\hat{\mathbf{Q}}$ is the following:

$$\hat{\mathbf{Q}} = \mathbf{X_w B} = \mathbf{X_w}\ [\mathbf{X'_w X_w}]^{-1}\mathbf{X'_w}\ \bar{\mathbf{Q}} \tag{C.4}$$

This is also equal to:

$$\hat{\mathbf{Q}} = \mathbf{D}(p_{i+})^{1/2}\,\mathbf{XB} = \mathbf{D}(p_{i+})^{1/2}\,\mathbf{X}\ [\mathbf{X}'\ \mathbf{D}(p_{i+})\ \mathbf{X}]^{-1}\mathbf{X}'\ \mathbf{D}(p_{i+})^{1/2}\,\bar{\mathbf{Q}} \tag{C.5}$$

2) - 3) As in step 2 of RDA, the covariance matrix $\mathbf{S} = \hat{\mathbf{Q}}'\hat{\mathbf{Q}}$ is computed (there is no division by degrees of freedom in CA and CCA), followed by eigenvalue decomposition (referred to as principal component analysis in step 3 of RDA) to reduce the dimensionality of the solution. CCA is thus a weighted form of RDA, approximating chi-square distances among the rows (sites) of matrix **Y**, subject to the constraint that the canonical axes are weighted linear combinations of the explanatory variables. In CCA, the number of canonical eigenvectors cannot exceed the minimum of $p - 1$, m, and $n - 1$.

4) - 5) - 6) Two main types of scaling, which may be applied to matrix **U** of the eigenvectors of **S**, are commonly used by biologists to draw biplot ordination diagrams when analyzing species presence-absence or abundance data. Other types of scaling are described by ter Braak (1987, 1990), ter Braak and Smilauer (1998) and Legendre and Legendre (1998).

• Scaling type 1 — Assuming that sites are rows and species are columns in **Y**, this scaling is the most appropriate if one is primarily interested in the ordination of sites: the sites (whose coordinates are found in matrix **F**, below) are plotted at the centroids of the species (whose coordinates are found in matrix **V**, below). In full-dimensional matrix **F**, distances among the sites are equal to the chi-square distances among the rows of matrix **Y**.

• Scaling type 2 — Assuming that sites are rows and species are columns in **Y**, this is the most appropriate scaling if one is primarily interested in the relationships among species: the species (whose coordinates are found in matrix $\hat{\mathbf{F}}$, below) are plotted at the centroids of the sites (whose coordinates are found in matrix $\hat{\mathbf{V}}$, below). In full-dimensional matrix $\hat{\mathbf{F}}$, distances among the species are equal to the chi-square distances among the columns of matrix **Y**.

The construction and interpretation of CCA biplots is discussed in more detail by ter Braak and Verdonschot (1995, Table 2), ter Braak and Smilauer (1998) and Legendre and Legendre (1998).

Matrix **V** of species scores (for scaling type 1) and matrix $\hat{\mathbf{V}}$ of site scores (for scaling type 2) are obtained from **U** using the weighs given by the diagonal matrices $\mathbf{D}(p_{+j})^{-1/2}$ and $\mathbf{D}(p_{i+})^{-1/2}$ containing inverses of the square roots of the column and row weights of **Y**, respectively:

$$\mathbf{V} = \mathbf{D}(p_{+j})^{-1/2}\mathbf{U} \qquad \text{(C.6)}$$

$$\hat{\mathbf{V}} = \mathbf{D}(p_{i+})^{-1/2}\,\bar{\mathbf{Q}}\mathbf{U}\Lambda^{-1/2} \qquad \text{(C.7)}$$

where $\Lambda^{-1/2}$ is a diagonal matrix containing inverses of the square roots of the eigenvalues of **S**. Matrices **F** (site scores for scaling type 1) and $\hat{\mathbf{F}}$ (species scores for scaling type 2) are found using:

$$\mathbf{F} = \hat{\mathbf{V}}\Lambda^{1/2} \qquad \text{(C.8)}$$

$$\hat{\mathbf{F}} = \mathbf{V}\Lambda^{1/2} \qquad \text{(C.9)}$$

The site scores which are weighted linear combinations of the explanatory variables, corresponding to Eq. B.6 of RDA (Appendix B), are derived from $\hat{\mathbf{Q}}$ as follows:

For scaling type 1: $$\mathbf{cord}_{(\text{space of explanatory variables } \mathbf{X})} = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{Q}}\mathbf{U} \qquad \text{(C.10)}$$

For scaling type 2: $$\mathbf{cord}_{(\text{space of explanatory variables } \mathbf{X})} = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{Q}}\mathbf{U}\Lambda^{-1/2} \qquad \text{(C.11)}$$

Weighted linear correlations between variables **x** of **X** and constrained ordination axes **cord** in space **X** (from Eq. C.10 or C.11) are used for representing the explanatory variables in biplots; the weight w_i associated with each row i of **cord** and **x** is p_{i+}. The weighted linear correlation $R_{\mathbf{cord,x}}$ is calculated as follows:

$$R_{\mathbf{cord,x}} = \frac{(\sum_{i=1}^{n} w_i cord_i x_i - ((\sum_{i=1}^{n} w_i x_i)(\sum_{i=1}^{n} w_i cord_i)/(\sum_{i=1}^{n} w_i)))}{\sqrt{((\sum_{i=1}^{n} w_i x_i^2) - ((\sum_{i=1}^{n} w_i x_i)^2 /(\sum_{i=1}^{n} w_i)))((\sum_{i=1}^{n} w_i cord_i^2) - ((\sum_{i=1}^{n} w_i cord_i)^2 /(\sum_{i=1}^{n} w_i)))}} \qquad \text{(C.12)}$$

With scaling type 2, the correlations are used directly as biplot scores for the environmental variables. With scaling 1, the correlations have to be weighted using Eq. B.7 before they are used as biplot scores.

Literature Cited

Legendre, P., and L. Legendre. 1998. Numerical ecology. Second English edition. Elsevier, Amsterdam, The Netherlands.

ter Braak, C. J. F. 1987. Ordination. Pages 91-173 in: R. H. G. Jongman, C. J. F. Ter Braak, and O. F. R. van Tongeren, editors. Data analysis in community and landscape ecology. Centre for Agricultural Publishing and Documentation, Wageningen, The Netherlands. Reissued in 1995 by Cambridge Univ. Press, Cambridge, England.

ter Braak, C. J. F. 1990. Update notes: CANOCO version 3.10. Agricultural Mathematics Group, Wageningen, The Netherlands.

ter Braak, C. J. F. and P. Smilauer. 1998. CANOCO reference manual and user’s guide to Canoco for Windows: software for canonical community ordination (version 4). Microcomputer Power, Ithaca, New York, USA.

ter Braak, C. J. F. and P. F. M. Verdonschot. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquatic Sciences **57**: 255-289.

APPENDIX D

ECOLOGICAL ARCHIVES E083-018-A4

TESTS OF SIGNIFICANCE IN POLYNOMIAL RDA AND CCA

Description of the permutation testing procedure

Permutation tests can be carried out in linear or polynomial RDA or CCA. The most general null hypothesis states that there is no special relationship between the response and explanatory variables (independence of **Y** and **X**), or that the model is not a significant representation of the response data. (Tests of significance for individual canonical eigenvalues are not discussed here; they are described in ter Braak and Smilauer [1998] and in Legendre and Legendre [1998]). RDA or CCA is computed as described in Fig. 1 of the main paper and the following pseudo-*F* statistic is computed for the unpermuted data:

$$\text{pseudo-}F = \frac{\text{variance of } \hat{\mathbf{Y}} \ (\text{or } \hat{\mathbf{Q}})}{\text{Total variance of } \mathbf{Y} \ (\text{or } \bar{\mathbf{Q}}) \ - \ \text{variance of } \hat{\mathbf{Y}} \ (\text{or } \hat{\mathbf{Q}})} \qquad \text{(D.1)}$$

We call this statistic a "pseudo-*F*" because the degrees of freedom are not included in the numerator and denominator. The full *F*-statistic customarily used in canonical analysis, which is also sometimes called "pseudo-*F*" but with a different meaning, contains degrees of freedom; it is described in the above-mentioned books. Degrees of freedom are multiplicative constants through the permutations; thus, including them, or not, does not change the outcome of the tests. They are not included here because the number of parameters in the polynomial model may change from permutation to permutation; this number is used in the formula for computing the degrees of freedom of a regular *F*-statistic. Simulations, reported in the next section, show that a permutation test of significance based upon the pseudo-*F* statistic described in Eq. D.1 has correct type I error.

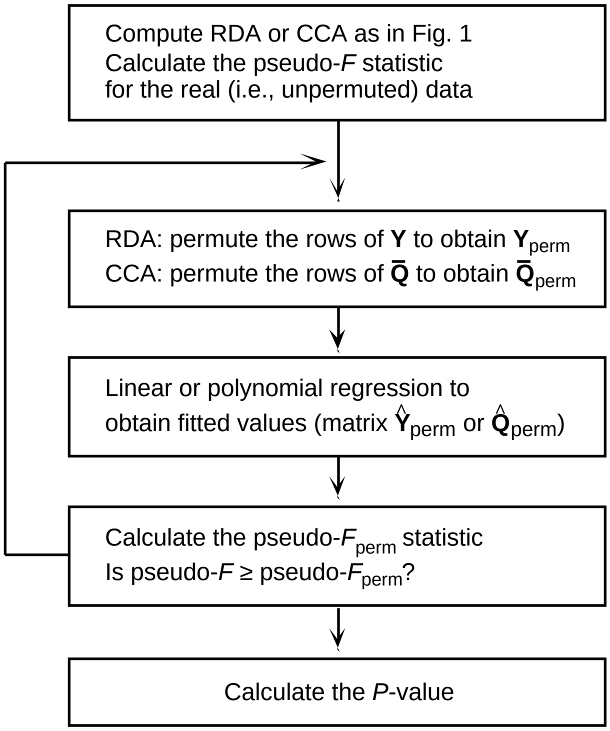
Compute RDA or CCA as in Fig. 1
Calculate the pseudo-F statistic
for the real (i.e., unpermuted) data
RDA: permute the rows of Y to obtain Y_perm
CCA: permute the rows of Q̄ to obtain Q̄_perm
Linear or polynomial regression to
obtain fitted values (matrix Ŷ_perm or Q̂_perm)
Calculate the pseudo-F_perm statistic
Is pseudo-F ≥ pseudo-F_perm?
Calculate the P-value

FIG. D.1. Permutation tests in linear and polynomial RDA and CCA.

To generate the null distribution (Fig. D.1), the rows of $\mathbf{Y}$ (or $\overline{\mathbf{Q}}$) are permuted at random to obtain the matrix $\mathbf{Y_{perm}}$ (or $\overline{\mathbf{Q}}_{\mathbf{perm}}$). Linear or polynomial regression is computed using the (unpermuted) matrix of explanatory variables $\mathbf{X}$ to obtain a matrix of fitted values $\hat{\mathbf{Y}}_{\mathbf{perm}}$ (or $\hat{\mathbf{Q}}_{\mathbf{perm}}$). The pseudo-$F_{\text{perm}}$ statistic can be directly estimated for the permuted data from the variances in $\mathbf{Y_{perm}}$ (or $\overline{\mathbf{Q}}_{\mathbf{perm}}$) and in $\hat{\mathbf{Y}}_{\mathbf{perm}}$ (or $\hat{\mathbf{Q}}_{\mathbf{perm}}$) using Eq. D.1. After repeating the permutation and calculation of the pseudo-F_{perm} statistic a large number of times, the probability (P-value) of the data under the null hypothesis is computed as the proportion of pseudo-F_{perm} values that are larger than or equal to pseudo-F. Following Hope (1968), the pseudo-F value obtained for the unpermuted data is included in the null distribution of pseudo-F_{perm} values. Generally, analyses providing P-values smaller than or equal to 0.05 are considered significant. A lower significance level should be used for many ecological problems in view of the fact that spatial autocorrelation is present most field ecological data sets (Legendre 1993).

Which model is the most appropriate to describe the data if the linear and polynomial models are both significant? To answer this question, we used a permutation procedure to assess the difference in variance accounted for between the polynomial model and the linear model nested into it. Essentially, the procedure is the following:

1) Polynomial RDA (or CCA) is used to find the variance of $\mathbf{Y}$ (or $\overline{\mathbf{Q}}$) accounted for by the polynomial model, $\text{Var}_{\text{polynomial}}$.

2) Linear RDA (or CCA) is used to find the variance of $\mathbf{Y}$ (or $\overline{\mathbf{Q}}$) accounted for by the linear model, $\text{Var}_{\text{linear}}$.

3) The variance of the difference between the two models is obtained by subtraction: $\text{Var}_{\text{polynomial}} - \text{Var}_{\text{linear}}$. Calculate the pseudo-$F$ statistic for this difference, using Eq. D.2:

$$\text{pseudo-}F = \frac{\text{Var}_{\text{polynomial}} - \text{Var}_{\text{linear}}}{\text{Total variance of } \mathbf{Y} (\text{or } \overline{\mathbf{Q}}) - \text{Var}_{\text{polynomial}}} \qquad \text{(D.2)}$$

This equation is constructed in the same way as that of an F-statistic for testing the significance of additional explanatory variables in nested regression models. Simulation, reported in the next section, show that a test of significance for the difference in explained variation between nested models, based upon this statistic, has correct type I error.

4) Permute matrix $\mathbf{Y}$ (or $\overline{\mathbf{Q}}$), to obtain matrix $\mathbf{Y}_{\text{perm}}$, (or $\overline{\mathbf{Q}}_{\text{perm}}$,). Repeat the calculations for the permuted matrix $\mathbf{Y}_{\text{perm}}$, (or $\overline{\mathbf{Q}}_{\text{perm}}$,), from which a pseudo-$F_{\text{perm}}$ statistic is obtained. Repeat this step a large number of times.

5) Test the significance of the pseudo-F statistic against the distribution of the pseudo-F_{perm}, as above, after incorporating the pseudo-F value into the distribution. The smaller the P-value for the difference between the two models, the more appropriate is the polynomial model.

Permutation of the rows of matrix $\mathbf{Y}$ (or $\overline{\mathbf{Q}}$), as performed here, is adequate in all instances where there are no covariables. For the test of the difference between the two models, an alternative would be to implement a procedure based upon permutation of the residuals of a reduced or a full regression model, as described in ter Braak and Smilauer (1998) and Legendre and Legendre (1998). These methods would procure an improvement in type I error, over the permutation of the rows of matrix $\mathbf{Y}$, only in cases where the covariable matrix contains extreme outliers (Anderson and Legendre 1999).

Do the Permutation Tests Have Correct Type I Error and Good Power?

This section aims at establishing that the probabilities obtained from the permutation tests described in the previous section have correct type I error (and thus provide valid tests of significance), and that they have good power, allowing to discriminate between the linear and polynomial relationships. To accomplish this, we ran a large number of numerical simulations using computer-generated data sets whose properties were known. Simulations are a standard approach in statistics because they allow verification of the properties of statistical procedures in situations where the answer is known (see, e.g., Milligan 1996).

First, we generated many data tables **Y** (response) and **X** (explanatory) containing random numbers. They represented situations where the null hypothesis of the test was true. So, a canonical analysis should not find these data sets to be related, except by chance; finding unrelated data sets to be significantly related is referred to as type I error. A test of significance is said to have correct rejection rate at significance level α (decided a priori by the user) if the null hypothesis is rejected in a proportion of the cases approximately equal to α, when using data generated to correspond to the null hypothesis. The test is said to be valid if the rejection rate is not larger than the significance level α, for any value of α, when the null hypothesis is true (Edgington 1995).

Two series of simulations were carried out, corresponding to two standard applications of RDA. In the first series of 1000 simulations, the data in **Y** were random numbers drawn from a multivariate normal distribution with variances of 1 and covariances of 0; in this series, **Y** simulated a matrix of standardized physical variables. In the second series of 1000 simulations, the data in **Y** were drawn at random from a standard lognormal distribution, to simulate species abundance data. In both series, the data in **X**, representing the explanatory variables, were multivariate random normal.

Matrices **X** and **Y** had the following parameters: *n*, the number of rows in **Y** and **X**, was 10; *p*, the number of columns in **Y**, was 5; *m*, the number of columns in **X**, was 5. For each data set, 499 random permutations of the rows of **Y** were done. We tallied the rates of rejection of the null hypothesis for 20 different values of α for standard linear RDA (using linear multiple regression), for polynomial RDA, and for the difference between the polynomial and linear solutions. We also computed confidence intervals for two commonly used values of significance level: $\alpha = 0.05$ and $\alpha = 0.10$.

Results of this study are presented in Tables D.1 and D.2. They show that in all cases, the null hypothesis (H_0) was rejected with frequencies very close to α; the 95% confidence intervals of the rejection rates always included the corresponding α values. We repeated the simulations with different values of matrix parameters *n*, *p* and *m* (*n* varying from 5 to 25, *p* from 5 to 15, and *m* from 5 to 10). The results, not reported in detail here, are very similar to those found in Tables 1 and 2. The simulation results confirm that the permutation tests described in the previous section have correct α-significance level. In other words, if no relationship exists between **Y** and **X**, the tests make type I errors at the rate predicted by the α-significance level.

Secondly, we designed and ran simulations for type II error, to determine if the test of significance for the difference between the two models (Eq. D.2) led to correct decisions.

1) We generated data sets where each of the **Y** variables ($n = 10$, $p = 5$, $m = 3$) was an independently-constructed *linear function* of the 3 variables in **X**, plus error. With normal error (results are not presented in detail here), the linear and polynomial RDA models were both significant, but the difference in explained variation between the polynomial and linear models was nearly never significant, as expected, when testing at $\alpha = 0.05$ or $\alpha = 0.10$. The few instances of significant differences represent type I error for the test of difference between the two models (Eq. D.2); since the null hypothesis of no difference was true in these

simulations, the test was expected to reject the null hypothesis in a fraction α of the simulated data sets. Similar results were obtained when using log-normal error.

2) We also generated data sets where each of the **Y** variables (n = 25, p = 10, m = 3) was an independently-constructed *polynomial function* of the 3 variables in **X**, plus error. The results were the opposite: the polynomial model was always significant; so was the difference between the polynomial and linear models, even when the linear model was also significant.

We conclude that the test of significance for the difference between the two models led to the correct decision in nearly all cases, finding the linear model to be the most appropriate when matrices **X** and **Y** had been generated to be linearly related, and the polynomial model to be the most appropriate when a polynomial relationship had been built between the two matrices.

LITERATURE CITED

Anderson, M. J., and P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation **62**: 271-303.

Edgington, E. S. 1995. Randomization tests. 3rd edition. Marcel Dekker, Inc., New York.

Hope, A. C. A. 1968. A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society **B 30**: 582-598.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology **74**: 1659-1673.

Legendre, P., and L. Legendre. 1998. Numerical ecology. Second English edition. Elsevier, Amsterdam, The Netherlands.

Milligan, G. W. 1996. Clustering validation – Results and implications for applied analyses. 341-375 in: P. Arabie, L. J. Hubert & G. De Soete [eds.] Clustering and Classification. World Scientific Publ. Co., River Edge, New Jersey.

ter Braak, C. J. F. 1994. Canonical community ordination. Part I: Basic theory and linear methods. Écoscience **1**: 127-140.

TABLE D.1. First set of simulations: random normal data in **Y** and random normal data in **X**. The upper portion of the table reports rejection rates of the null hypothesis at the significance levels found in the left column, after studying 1000 pairs of random data sets; 499 permutations were used for each test. The lower portion reports 95% confidence intervals of the rejection rates for two widely used significance levels, α = 0.05 and α = 0.10.

	Rate of rejection of H_0 for		
Significance level α	Linear regression (L)	Polynomial regression (P)	Difference in explained variance between polynomial and linear models (P-L)
0.05	0.050	0.047	0.046
0.10	0.098	0.086	0.098
0.15	0.149	0.156	0.146
0.20	0.196	0.195	0.197
0.25	0.246	0.247	0.250
0.30	0.306	0.295	0.294
0.35	0.349	0.335	0.349
0.40	0.401	0.383	0.401
0.45	0.450	0.441	0.449
0.50	0.502	0.491	0.493
0.55	0.553	0.546	0.546
0.60	0.601	0.594	0.594
0.65	0.637	0.641	0.646
0.70	0.692	0.702	0.697
0.75	0.731	0.754	0.747
0.80	0.767	0.803	0.795
0.85	0.826	0.851	0.853
0.90	0.884	0.905	0.903
0.95	0.949	0.951	0.953
1.00	1.000	1.000	1.000

Significance	Linear regression		Polynomial regression		Difference P-L	
level α	C1	C2	C1	C2	C1	C2
0.05	0.0365	0.0635	0.0339	0.0601	0.0330	0.0590
0.10	0.0825	0.1135	0.0714	0.1006	0.0825	0.1135

Abbreviations: C1 and C2 are the limits of the 95% confidence intervals, for α = 0.05 and 0.10, computed for the rejection rates found after analyzing 1000 random data sets. Type 1 error of a test is correct if the confidence interval of the empirical rejection rate includes α.

TABLE D.2. Second set of simulations: random lognormal data in **Y** and random normal data in **X**. The upper portion of the table reports rejection rates of the null hypothesis at the significance levels found in the left column, after studying 1000 pairs of random data sets; 499 permutations were used for each test. The lower portion reports 95% confidence intervals of the rejection rates for two widely used significance levels, α = 0.05 and α = 0.10. Abbreviations as in Table D.1.

	Rate of rejection of H_0 for		
Significance level α	Linear regression (L)	Polynomial regression (P)	Difference in explained variance between polynomial and linear models (P-L)
0.05	0.045	0.044	0.039
0.10	0.088	0.092	0.091
0.15	0.136	0.137	0.135
0.20	0.195	0.188	0.188
0.25	0.237	0.231	0.251
0.30	0.279	0.269	0.306
0.35	0.326	0.332	0.356
0.40	0.360	0.393	0.396
0.45	0.408	0.442	0.453
0.50	0.445	0.496	0.500
0.55	0.511	0.543	0.543
0.60	0.566	0.592	0.604
0.65	0.623	0.647	0.650
0.70	0.690	0.697	0.698
0.75	0.738	0.746	0.749
0.80	0.782	0.794	0.798
0.85	0.840	0.847	0.853
0.90	0.885	0.895	0.908
0.95	0.953	0.950	0.962
1.00	1.000	1.000	1.000

Significance	Linear regression		Polynomial regression		Difference P-L	
level α	C1	C2	C1	C2	C1	C2
0.05	0.0321	0.0579	0.0313	0.0567	0.0270	0.0510
0.10	0.0733	0.1028	0.0770	0.1071	0.0760	0.1060

APPENDIX E

ECOLOGICAL ARCHIVES E083-018-A5

RESULTS OF POLYNOMIAL CCA OF THE SPIDER SPECIES DATA (SELECTED OUTPUT)

	Canonical axes					
Data type	I	II	III	IV	V	VI
Eigenvalues (with respect to total variance in $\overline{\mathbf{Q}}$ = 1.92296)						
	0.62894	0.37511	0.24337	0.17653	0.04974	0.02939
Fraction of total variance in $\overline{\mathbf{Q}}$						
	32.70698	19.50712	12.65596	9.17995	2.58668	1.52812
Cumulative fraction of total variance in $\overline{\mathbf{Q}}$ accounted for by axes I to VI						
	32.70698	52.21410	64.87006	74.05001	76.63669	78.16481
Species scores from Eq. C.9, matrix $\hat{\mathbf{F}}$ (scaling type 2)						
Al. accentuata	1.55036	–0.14821	0.26633	0.81398	0.40180	–0.30755
Al. cuneata	–0.17992	–0.19077	–0.00205	–0.12488	0.14257	–0.40530
Al. fabrilis	2.52533	0.95247	0.21864	1.05935	–0.42174	0.37730
Ar. lutetiana	–0.53087	0.25503	–0.44826	0.10140	–0.33373	0.19051
Ar. perita	3.91140	2.10706	–1.24542	–2.67197	0.11346	–0.11043
Au. albimana	–0.38288	0.14117	–0.41733	0.12523	0.20572	0.16340
Pa. lugubris	–0.32128	0.61698	2.00915	–0.52270	0.39451	0.25796
Pa. monticola	0.53335	–1.36366	0.05890	–0.28172	–0.08014	0.10428
Pa. nigriceps	–0.45556	0.20890	–0.45936	0.09609	0.16259	0.19589
Pa. pullata	–0.36954	0.11816	–0.37024	0.07745	0.13450	0.02186
Tr. terricola	–0.31803	0.22305	0.15007	–0.05952	–0.14189	–0.08913
Zo. spinimana	–0.41098	0.32522	0.16851	0.06653	–0.49435	–0.08025
Site scores from Eq. C.7, matrix $\hat{\mathbf{V}}$ (scaling type 2)						
Site 1	0.19190	–0.83192	–0.09555	0.13373	0.48920	–1.22681
Site 2	–0.55696	0.43411	–0.51925	0.15622	0.05135	0.21341
Site 3	–0.07859	–0.25295	–0.23664	0.18236	0.60622	–1.50035
Site 4	–0.51493	0.33292	–0.58710	0.19071	–0.09421	–0.13048
Site 5	–0.58522	0.40912	–0.89278	0.23397	0.86853	1.06013
Site 6	–0.49729	0.26545	–0.25896	0.03908	–2.16265	–0.47017
Site 7	–0.47088	0.14143	–0.70784	0.10572	0.13999	0.95180
Site 8	–0.47263	0.88678	4.11220	–1.56466	3.47613	2.48576
Site 9	0.68911	–2.95056	0.15118	–1.14120	–0.97201	2.02230

Site 10	1.09945	–2.92298	0.36349	–0.55441	–0.84824	1.94692
Site 11	1.03033	–2.92079	0.32830	–0.57574	–0.43813	1.41973
Site 12	0.71979	–2.42159	0.22390	–0.58130	–0.01566	–0.36876
Site 13	–0.40555	0.05911	–0.46141	0.01814	0.11136	–1.19280
Site 14	–0.48114	0.26181	–0.17299	–0.01462	–1.38733	–0.74329
Site 15	–0.51500	0.77094	1.95994	–0.71338	–1.58809	–1.06710
Site 16	–0.51036	0.65473	1.51807	–0.59788	–1.33662	–1.80145
Site 17	–0.51238	0.65955	1.17666	–0.45481	–2.59198	–2.28087
Site 18	–0.49480	0.69538	1.81713	–0.74849	–0.94208	–1.62016
Site 19	–0.51398	0.87640	2.88123	–1.05265	0.21075	0.45227
Site 20	–0.49909	0.86903	3.12334	–1.17034	0.65863	0.22280
Site 21	–0.47251	0.86246	3.66255	–1.37506	1.51583	1.67290
Site 22	3.59010	1.75540	0.07743	2.16450	–2.54873	3.89519
Site 23	3.11351	1.06618	0.29769	0.99498	0.60038	0.35883
Site 24	3.62149	1.59815	0.12049	2.33853	–1.91440	3.27871
Site 25	0.81245	–0.04527	0.44939	1.26706	–1.45461	0.53678
Site 26	5.10899	4.10097	–2.96940	–7.93872	0.39989	–1.03919
Site 27	3.14631	0.75153	0.46391	3.08143	3.28692	–4.00636
Site 28	2.72787	0.35159	0.68754	3.38685	–0.29284	0.75589

Biplot scores of environmental variables, from Eq. 12

Water	–0.92432	0.18650	–0.23523	–0.18804	–0.37686	–0.31522
Reflection	0.84674	–0.60030	–0.32579	–0.53371	0.21990	–0.28005
Calamagrostis	–0.52146	0.22616	–0.70727	0.14786	0.30211	0.52232
Coryneporus	0.78682	–0.68336	0.08988	–0.17240	0.05986	0.17525

Biplot scores of environmental variables: weighted correlations

Water	–0.66479	0.17312	–0.23287	–0.04559	–0.33191	–0.19449
Reflection	0.76280	–0.28002	–0.28547	–0.03780	0.14508	–0.27962
Calamagrostis	–0.49123	0.22202	–0.69103	0.14315	0.28329	0.16033
Corynephorus	0.56135	–0.66456	0.08976	–0.00597	0.04519	0.04974
Water^2	0.60755	0.07809	–0.04507	0.17987	–0.19515	–0.25764
Reflection^2	0.73372	0.28057	–0.03136	–0.46095	–0.05465	–0.14486
Calamagrostis^2	0.18803	0.03716	0.16909	0.03316	–0.11249	0.49264
Corynephorus^2	0.32062	–0.67672	0.08182	–0.06862	0.02769	–0.01521
Water x Refl.	–0.79020	–0.10389	0.14754	0.10273	0.22799	0.03177
Water x Calam.	0.60467	–0.16110	–0.04745	0.05545	0.42990	0.10455
Water x Corynep.	–0.54738	0.52220	–0.02685	–0.12139	0.01754	0.06717
Refl. x Calam.	–0.65178	0.14440	0.52247	–0.00110	–0.31599	–0.29830
Refl. x Corynep.	0.50857	–0.46226	0.17410	–0.05155	–0.00590	0.17733
Calam. x Corynep.	–0.45070	0.66288	0.15644	–0.04998	–0.15946	–0.11593

Notes: Matrix **Y**: hunting spider species 1 to 12. Matrix **X**: water content, reflection of soil surface, percent cover by *Calamagrostis epigejos*, percent cover by *Corynephorus canescens*. Either set of biplot scores can be used to represent the environmental variables in biplots.

Erratum

Page 1158 column 2: the equation for x_{i23} should read

$$x_{i23} = 0.1440 - 0.0030x_{i3} + 0.0075x_{i2} - 0.000034x_{i3}^2 - 0.000071x_{i2}x_{i3} - 0.000390x_{i2}^2$$