Co-correspondence analysis: a new ordination method to relate two community compositions

ter Braak, C. J. F., & Schaffers, A. P.

Co-correspondence analysis: a new ordination method to relate two community compositions

Cajo J.F. ter Braak[1,3]& André P. Schaffers[2]

[1]Biometris, Wageningen University and Research Centre, PO Box 100, 6700 AC Wageningen, The Netherlands

[2]Nature Conservation and Plant Ecology Group, Department of Environmental Sciences, Wageningen University and Research Centre, Bornsesteeg 69, 6708 PD Wageningen, The Netherlands

[3]Corresponding author: cajo.terBraak@wur.nl

Abstract

A new ordination method, called co-correspondence analysis, is developed to relate two types of communities (*e.g.* a plant community and an animal community) sampled at a common set of sites in a direct way. The method improves the simple, indirect approach of applying correspondence analysis (reciprocal averaging) to the separate species data sets and correlating the resulting ordination axes. Co-correspondence analysis maximizes the weighted covariance between weighted averaged species scores of one community and weighted averaged species scores of the other community. It thus attempts to identify the patterns that are common to both communities. Both a symmetric descriptive and an asymmetric predictive form are developed. The symmetric form relates to co-inertia analysis and the asymmetric, predictive form to partial least squares regression. In two examples the predictive power of co-correspondence analysis is compared with that of canonical correspondence analyses on syntaxonomic and environmental data. In the first example carabid beetles in roadside verges are shown to be more closely related to plant species composition than to vegetation structure (biomass, height, roughness, among others), and in the second example bryophytes in spring meadows are shown to be more closely related to the species composition of the vascular plants than to the measured water chemistry.

Key words: community data; correspondence analysis; canonical correspondence analysis; co-inertia analysis; gradient analysis; ordination; partial least squares; species-environment relationships

Introduction

Are carabid beetles in roadside verges more closely related to vegetation structure than to plant species composition or is the reverse true? To answer such a question from field data we need quantitative methods, one to predict the beetle community (that is, the incidences and abundances of each of the beetle species) from a set of variables characterizing vegetation structure (biomass, height, roughness, among others) and another to predict the beetle community from the plant community. (The word 'community' is used here to mean an assemblage of taxa.) The plant community can be thought of as a bio-assay of the environment (Persson 1981, Prentice and Cramer 1990). A related question is then: is the beetle community better predicted from the measured environmental variables than from the plant community? The data on beetles, plants, vegetation structure and environment that we consider are all from a common set of sites.

For relating a biological community to a set of (environmental) predictor variables, we have at our disposal methods of multivariate direct gradient analysis (ter Braak and Prentice 1988): canonical correspondence analysis for community data showing unimodal responses (ter Braak 1986) and redundancy analysis for data with linear responses (Jongman et al. 1995, Legendre and Legendre 1998). These methods are part of the computer program CANOCO (ter Braak and Šmilauer 2002). Canonical correspondence analysis (Table 1) is the method of choice (ter Braak and Prentice 1988, Jongman et al. 1995) when the community data show unimodal responses, have a strong qualitative (presence/absence) nature, and/or are sum-constrained (which means that the total abundance per site is fixed or determined by the sampling method). This method works to predict the beetle community from the variables on the vegetation structure or the environment but not to predict the beetle community from the plant community data. The reasons are two-fold: (1) canonical correspondence analysis breaks down when the number of predictor variables, the individual plant species in our case, is larger than the number of sites, and (2) the method uses linear combinations -weighted sums (Jongman et al. 1995) - of the predictor variables, which does not work well when the community data in the predictor role show a unimodal structure, a strong qualitative nature and/or are sum-constrained, *i.e.* when the predictor community is better analyzed by a correspondence analysis technique than by a linear technique such as principal components analysis (ter Braak and Juggins 1993, ter Braak et al. 1993, Frisvad and Norsker 1996).

A possible approach is to analyze the community data sets separately by (detrended) correspondence analysis (or another suitable indirect gradient analysis method) and correlate the first few axes from these analyses by calculating pairwise correlations (Hájek et al. 2002). This method is correlative instead of predictive. It can be made predictive by applying a (detrended) correspondence analysis solely to the community data in the predictor role and taking the first few resulting axes as predictor variables in a canonical correspondence analysis of the community data in the response role. This is a two-step (indirect) approach, which works fine if the major pattern of variation in the predictor community is important for the response community. This

need not to be the case. We therefore need a more generally applicable one-step (direct) approach that integrates the two steps of the analysis. With a one-step method the most important relationships are expressed in the first few axes so that one can be sure one does not miss something important. The new method need to integrate two correspondence analyses in a predictive fashion. Before we describe our new method, it is instructive to briefly discuss two related methodologies.

Co-inertia analysis (Dolédec and Chessel 1994) is a general eigenvector framework to relate two data sets in a symmetric way, *i.e.* neither set takes the response or predictor role. The ordination axes of co-inertia analysis maximize covariance. The strong points of co-inertia analysis are that it can deal with large numbers of variables in either set and that it includes, by way of pre-processing and row and column weighing, both methods related to principal components analysis (linear response) and methods related to correspondence analysis (unimodal response). To the best of our knowledge co-inertia analysis has, however, never been used to relate two correspondence analysis tables. The reason perhaps is that it is not immediately clear how to carry out such a co-inertia analysis because each correspondence analysis implies its own site weights (the site's total abundance of the species in the analysis) whereas co-inertia analysis requires common site weights. In this paper we overcome this difficulty and we show how to relate two correspondence analysis tables in the co-inertia framework.

Partial least squares (PLS) (Martens and Naes 1992) is a methodology for multivariate linear regression, popular in chemometrics, that can handle large numbers of predictor variables without much loss of predictive power. As in co-inertia analysis the ordination axes of PLS maximize covariance, but the constraints used in the maximization differ. We consider PLS the predictive counterpart of unweighted co-inertia analysis (de Jong and ter Braak 1994). ter Braak and co-workers proposed PLS-extensions of existing ecological methods (Table 1). The PLS-extension of canonical correspondence analysis, called CCA-PLS, can handle large numbers of predictor variables (ter Braak and Verdonschot 1995). CCA-PLS maximizes the covariance between weighted averaged species scores and linear combinations of the predictor variables. Because CCA-PLS takes linear combinations of the predictor variables, it is not the method we are looking for. Weighted averaging calibration (Table 1) predicts an environmental variable from community data by averaging indicator values of the species present a site (*e.g.* Persson 1981, Jongman et al. 1995). When the indicator values (the species scores of this method) for a particular environmental variable are unknown, they can be estimated from training data by weighted averaging regression (ter Braak and van Dam 1989, Birks et al. 1990, Fritz et al. 1991). The PLS-extension of weighted averaging regression and calibration (WA-PLS, ter Braak and Juggins 1993, ter Braak et al. 1993) uses ideas from PLS to estimate the indicator values. WA-PLS is popular in palaeoecology for reconstructing palaeoenvironments from fossil assemblages (Birks 1998). WA-PLS is called correspondence analysis partial least squares by Frisvad and Norsker (1996). The community data take the role of response variables in CCA-PLS and the role of predictor variables in WA-PLS (Table 1). These methods can be applied by using

existing computer programs for PLS (Martens 1986, Wise and Gallagher 2000, Wold 2002) by pre-processing the community data before it enters PLS and by post-processing the results of PLS (ter Braak et al. 1993, ter Braak and Verdonschot 1995). It can thus be expected that the new method proposed in this paper can formally be expressed as a PLS with particular pre- and post-processing steps. The pre- and post-processing steps are, however, not simply those applied to the community data in CCA-PLS and WA-PLS: we must first solve the problem of differences in implied site weights.

Here we propose a new ordination method, called co-correspondence analysis, to relate one community data set to another in a direct way (Table 1). Our starting point is the reciprocal averaging approach to correspondence analysis (Hill 1973, ter Braak and Prentice 1988). We resolve the problem of different sets of site weights, noted above, by posing an explicit maximization criterion for co-correspondence analysis: the (possibly weighted) covariance between weighted averaged species scores of one community with weighted averaged species scores of the other community. The weighted averages in the maximization criterion replace the linear combinations used in co-inertia analysis, PLS, and principal component analysis. This replacement makes the new method a correspondence analysis type of method, suited for unimodal data. We define a symmetric, descriptive form, which fits in the framework of co-inertia analysis, and an asymmetric, predictive form, which is a weighted form of PLS. Predictive co-correspondence analysis relates to correspondence analysis just as non-weighted PLS relates to principal component analysis.

We demonstrate the use of co-correspondence analysis in two examples and compare its predictive power with that of canonical correspondence analysis on, among others, environmental data. In the first example we show that carabid beetles in roadside verges are more closely related to plant species composition than to vegetation structure or environmental data, and in the second that bryophytes in spring meadows are more closely related to the species composition of the vascular plants than to the measured water chemistry.

Theory

Notation and computation

Let $\mathbf{Y}_1 = \{y_{1ik}\}[i = 1 \ldots n; k = 1 \ldots p]$ and $\mathbf{Y}_2 = \{y_{2il}\}[i = 1 \ldots n; l = 1 \ldots q]$ be $n \times p$ and $n \times q$ matrices containing the abundances of each of p and q species of community 1 and 2 at each of n sites, respectively. In the theory and examples that follow later on, one data set ($\mathbf{Y}_1$) is designated as the community in the "response role" and the other ($\mathbf{Y}_2$) as the community in the "predictor role". We then call the species in $\mathbf{Y}_1$ and $\mathbf{Y}_2$ response species and predictor species, respectively. By denoting summation across an index by a +, the site (row) weights are y_{1i+} and y_{2i+} and the species weights are y_{1+k} and y_{2+l}. These weights are also collected in the diagonal matrices $\mathbf{R}_1 = \text{diag}(\{y_{1i+}\})$, $\mathbf{R}_2 = \text{diag}(\{y_{2i+}\})$, $\mathbf{K}_1 = \text{diag}(\{y_{1+k}\})$ and $\mathbf{K}_2 = \text{diag}(\{y_{2+l}\})$. To keep the formulae as simple as possible, we preprocess each abundance table by dividing its values by the grand total, so that $y_{1++} = 1$ and $y_{2++} = 1$. This does

not change the results of the analyses in this paper.

In addition, $\mathbf{R}_0$ will contain user-defined site weights in the combined analysis of $\mathbf{Y}_1$ and $\mathbf{Y}_2$. In predictive co-correspondence analysis, a logical choice is to make $\mathbf{R}_0$ equal to $\mathbf{R}_1$, whereas in the symmetric analysis, one can make $\mathbf{R}_0 = (\mathbf{R}_1 + \mathbf{R}_2)/2$. These choices are justified in the Discussion. In the general case, $\mathbf{R}_0 = \text{diag}(\{r_{i0}\})$ with $r_{i0} > 0$ and $r_{+0} = 1$. The results of the analysis are sets of scores for species and sites. We denote the vectors of species scores for the two sets by $\mathbf{u}_1$ and $\mathbf{u}_2$, and the vectors of site scores derived from these by $\mathbf{x}_1$ and $\mathbf{x}_2$. The superscript T denotes the transpose of a vector or matrix, $\mathbf{I}_p$ the $p \times p$ identity matrix and $\mathbf{1}_p$ a column vector of p ones.

Computations were carried out in MATLAB (MATLAB 2000) using the PLS_Toolbox version 2.1 (Wise and Gallagher 2000) and in Canoco for Windows 4.5 (ter Braak and Šmilauer 2002). The extra MATLAB functions for the methods presented in this paper and the example data are available as Ecological Archive.

Co-correspondence analysis: definition

Correspondence analysis of community data is a method that assigns scores to species and sites that have certain optimality properties. Weighted averaging is one of the key concepts herein (Jongman et al. 1995). When looking for optimal species scores, we derive site scores from the species scores by the method of weighted averaging. Correspondence analysis assigns species scores so as to maximize the variance of these site scores under the constraint that the assigned species scores have unit variance. (The method is symmetrical in that 'species' and 'site' can be interchanged in the optimization criterion). For mathematical reasons, the variances are weighted with each site and each species receiving a weight proportional to its abundance total.

When the species can be subdivided in two sets (for example, beetles and plants) and the object is to relate these two sets, there are other ways to assign species scores, simply because we can calculate two sets of site scores, one from each species set. Instead of aiming at maximum variance we can now aim at maximum covariance between the sets of site scores, as is done in co-inertia analysis (Dolédec and Chessel 1994) and in PLS (Martens and Naes 1992, ter Braak and de Jong 1998). More precisely, co-correspondence analysis seeks vectors of species scores $\mathbf{u}_1$ and $\mathbf{u}_2$ that

(1) maximize $\mathbf{x}_1^T \mathbf{R}_0 \mathbf{x}_2$ with $\mathbf{x}_1 = \mathbf{R}_1^{-1}\mathbf{Y}_1\mathbf{u}_1$ and $\mathbf{x}_2 = \mathbf{R}_2^{-1}\mathbf{Y}_2\mathbf{u}_2$,

with $\mathbf{R}_0$ an $n \times n$ diagonal matrix with user-defined weights for the sites, and

subject to the constraints that the species scores have zero mean and unit variance:

(2) $\mathbf{1}_p^T \mathbf{K}_1 \mathbf{u}_1 = 0$, $\mathbf{1}_q^T \mathbf{K}_2 \mathbf{u}_2 = 0$ and $\mathbf{u}_1^T \mathbf{K}_1 \mathbf{u}_1 = 1$, $\mathbf{u}_2^T \mathbf{K}_2 \mathbf{u}_2 = 1$.

In words, what is maximized $(\mathbf{x}_1^T\ \mathbf{R}_0\ \mathbf{x}_2)$ is the covariance between the two sets of site scores with common site weights; the covariance is maximized by finding the appropriate vectors of species scores $\mathbf{u}_1$ and $\mathbf{u}_2$. Note however that what is maximized is formally not a covariance, because it is not guaranteed that the site scores $\mathbf{x}_1$ and $\mathbf{x}_2$ have zero mean. Yet, their mean will be close to zero if $\mathbf{R}_0 \approx \mathbf{R}_1 \approx \mathbf{R}_2$; and if, for example, $\mathbf{R}_0 = \mathbf{R}_1$, then the $\mathbf{R}_0$–weighted mean of $\mathbf{x}_1$ is zero and adding a constant to $\mathbf{x}_2$ would not change the criterion. Equations (1) and (2) are chosen to give a close link with standard correspondence analysis. Specifically, if $\mathbf{Y}_1 = \mathbf{Y}_2$ and we choose $\mathbf{R}_0 = \mathbf{R}_1$, then the Eqs (1) and (2) specify a correspondence analysis. By applying the Lagrange multiplier method (Magnus and Neudecker 1988) as in ter Braak and de Jong (1998), the maximization problem (1) - (2) leads to the transition formulae of co-correspondence analysis (with λ the Lagrange multiplier, which turns out to be an eigenvalue):

(3) $$\lambda \mathbf{u}_1 = \mathbf{K}_1^{-1}\mathbf{Y}_1^T\mathbf{x}_2^* \text{ with } \mathbf{x}_2^* = \mathbf{R}_1^{-1}\mathbf{R}_0\mathbf{x}_2$$

(4) $$\mathbf{x}_1 = \mathbf{R}_1^{-1}\mathbf{Y}_1(\mathbf{u}_1 - \bar{\mathbf{u}}_1) \text{ with } \bar{\mathbf{u}}_1 = y_{1++}^{-1}\sum_{k=1}^{p} y_{1+k}\mathbf{u}_1$$

(5) $$\mathbf{u}_2 = \mathbf{K}_2^{-1}\mathbf{Y}_2^T\mathbf{x}_1^* \text{ with } \mathbf{x}_1^* = \mathbf{R}_2^{-1}\mathbf{R}_0\mathbf{x}_1$$

(6) $$\mathbf{x}_2 = \mathbf{R}_2^{-1}\mathbf{Y}_2(\mathbf{u}_2 - \bar{\mathbf{u}}_2) \text{ with } \bar{\mathbf{u}}_2 = y_{2++}^{-1}\sum_{l=1}^{q} y_{2+l}\mathbf{u}_2 \ .$$

The required optimal species scores are the centered ones, $\mathbf{u}_1 - \bar{\mathbf{u}}_1$ and $\mathbf{u}_2 - \bar{\mathbf{u}}_2$. Apart from terms like $\mathbf{R}_1^{-1}\mathbf{R}_0$ which account for differences in weights among the sets, all equations are weighted averages:

- the species scores of one set are obtained as weighted averages of the other set's site scores (Eqs (3) and (5)) and
- the site scores are weighted averages of the species scores of their own set, as required by Eq. (1).

By comparison, the transition formulae of canonical correspondence analysis (ter Braak 1986) mix weighted averaging equations with equations from multiple regression.

As in correspondence analysis (Jongman et al. 1995) and canonical correspondence analysis (ter Braak 1986), the transition formulae can be solved by an iteration algorithm, which starts with arbitrary site scores and then applies the transition

formulae in turn (ignoring λ). In contrast with the iteration algorithm of canonical correspondence analysis (ter Braak 1986), centering and standardization should be applied to each set of species scores, and not to the site scores, as is evident from Eq. (2) and Eqs (4) and (6). The MATLAB function COCA-trans in the digital supplement gives all details. The algorithm yields the scores for the first ordination axis of co-correspondence analysis.

The transition equations can be condensed to an eigenvalue problem. Solving for the dominant eigenvalue λ yields the scores for the first ordination axis. The maximum covariance in Eq. (1) is the square root of λ. In the next two subsections we consider two different ways of extracting further ordination axes. In the first way the analysis is fully symmetrical in $\mathbf{Y}_1$ and $\mathbf{Y}_2$, whereas it is asymmetric in the second, yielding symmetric and predictive co-correspondence analysis, respectively. Despite the theoretical differences, the results of two methods need not be very different in practice (Burnham et al. 1996).

Symmetric co-correspondence analysis as a form of co-inertia analysis

In symmetric co-correspondence analysis the eigenvalue problem of the previous subsection is solved not only for the first, dominant eigenvalue, but also for subdominant eigenvalues, giving rise to further ordination axes. The resulting species scores of different axes are orthonormal, *i.e.* $\mathbf{U}_1^T\mathbf{K}_1\mathbf{U}_1 = \mathbf{I}_A$ and $\mathbf{U}_2^T\mathbf{K}_2\mathbf{U}_2 = \mathbf{I}_A$ with A axes arranged in columns (A = min(n, p, q)-1; see ter Braak and de Jong 1998). This analysis fits in the framework of co-inertia analysis (Dolédec and Chessel 1994) for relating two data tables, each represented by a statistical triplet consisting of the data table itself and column and row weights. Together the two statistical triplets define an eigenvalue problem. By straightforward (but tedious) algebra, it can be shown that symmetric co-correspondence analysis is the co-inertia analysis of the statistical triplets ($\mathbf{Q}_1$, $\mathbf{K}_1$, $\mathbf{R}_0$) and ($\mathbf{Q}_2$, $\mathbf{K}_2$, $\mathbf{R}_0$) with

$$(7) \qquad \mathbf{Q}_1 = \mathbf{R}_1^{-1}\mathbf{Y}_1\mathbf{K}_1^{-1} - \mathbf{1}_n\mathbf{1}_p^T \text{ and } \mathbf{Q}_2 = \mathbf{R}_2^{-1}\mathbf{Y}_2\mathbf{K}_2^{-1} - \mathbf{1}_n\mathbf{1}_q^T .$$

Each $\mathbf{Q}_s$ simply contains the residuals (o-e)/e with o = the observed abundance and e = (row total × column total) / (grand total), the expected abundance under row-column independence in the original abundance table $\mathbf{Y}_s$ (s=1,2), when treated as a contingency table. Therefore symmetric co-correspondence analysis can be carried out by any computer program that can perform co-inertia analysis, such as ADE-4 (Thioulouse et al. 1995) as outlined in Appendix A. To the best of our knowledge this form of co-inertia analysis has not been used before.

For comparison, separate correspondence analyses are based on the statistical triplets ($\mathbf{Q}_1$, $\mathbf{K}_1$, $\mathbf{R}_1$) and ($\mathbf{Q}_2$, $\mathbf{K}_2$, $\mathbf{R}_2$). These triplets cannot be used in a co-inertia analysis because of the differences in site weights between the sets if $\mathbf{R}_1 \neq \mathbf{R}_2$. The analysis of the previous section shows that the weighted averaging properties of correspondence analysis can be retained by replacing $\mathbf{R}_1$ and $\mathbf{R}_2$ by a single set of weights $\mathbf{R}_0$, which can be chosen by the user. This is not the first paper to combine a double-centered

table, such as $\mathbf{Q}_1$ in Eq. (7), with non-standard row and column weights. Thioulouse et al. (1995) did so in their multivariate analysis of local and global spatial patterns. Their analysis derives from maximization of the local (or global) spatial autocovariance, which is - like our criterion (1)- not a real covariance.

For further comparison, the simplest form of co-inertia, based on two principal components analyses, uses the triplets ($\mathbf{Y}_1$, $\mathbf{I}_p$, $\mathbf{I}_n$) and ($\mathbf{Y}_2$, $\mathbf{I}_q$, $\mathbf{I}_n$) and equals interbattery factor analysis (Tucker 1958, ter Braak 1990). It leads to transition formulae in which, compared to Eqs (3) - (6), weighted averages are replaced by weighted sums (with centering of site scores rather than species scores). Also for comparison, canonical correspondence analysis is formally the co-inertia of the triplets ($\mathbf{Q}_1$, $\mathbf{K}_1$, $\mathbf{R}_1$) and ($\mathbf{Z}$, $(\mathbf{Z}^T\mathbf{R}_1\mathbf{Z})^{-1}$, $\mathbf{R}_1$) with $\mathbf{Z}$ being the matrix containing the environmental data. To facilitate the comparison of the variance explained by co-correspondence analysis with that of a canonical correspondence analysis of $\mathbf{Y}_1$ with respect to $\mathbf{Z}$ we choose in the examples $\mathbf{R}_0 = \mathbf{R}_1$, although this destroys (strictly speaking) the symmetry of the method. For further arguments on this choice, see the Discussion section.

Predictive co-correspondence analysis as a form of partial least squares (PLS)

Partial least squares regression (PLS) works with ordination axes (called components or latent variables in this context) that maximize the covariance between linear combinations of response variables and of predictor variables, subject to particular constraints (Martens and Naes 1992). PLS differs from co-inertia analysis in two ways. First, PLS is less general than co-inertia analysis in that it does not use row and column weights. Second, the constraints used in PLS for the second and higher axes differ from those used in co-inertia analysis. Rather than requiring that the species scores be uncorrelated to those of previous axes, PLS requires that the site scores derived from the predictor variables be uncorrelated with the previously derived site scores. (There are actually two slightly different versions of multivariate PLS, NIPALS and SIMPLS (ter Braak and de Jong 1998) which need not concern us here). This simple difference from co-inertia analysis makes PLS asymmetric in the two data sets; PLS is a true regression method, and its main use is predictive: to predict one set of variables from the other set. In contrast, co-inertia analysis is symmetric in the two sets.

The question arises whether we might find a way to define a predictive version of co-correspondence analysis by placing it in the PLS framework instead of in the co-inertia framework. For this we must take account of the row and column weights. But there is a general mathematical trick to carry out a weighted analysis by a computer program that does not use weights (Seber 1977). The trick is to multiply each row of the data matrix by the square root of the row weight and, similarly, each column of the data matrix by the square root of the column weight. By applying this trick we obtain that predictive co-correspondence analysis of $\mathbf{Y}_1$ with respect to $\mathbf{Y}_2$ is PLS of the response matrix $\mathbf{Y}$ with respect to the predictor matrix $\mathbf{X}$ with

(8) $\mathbf{Y} = \mathbf{R}_0^{1/2}\mathbf{Q}_1\mathbf{K}_1^{1/2}$ and $\mathbf{X} = \mathbf{R}_0^{1/2}\mathbf{Q}_2\mathbf{K}_2^{1/2}$,

in which $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are derived from the original data $\mathbf{Y}_1$ and $\mathbf{Y}_2$ by Eq. (7). If $\mathbf{R}_0 = \mathbf{R}_1$, then $\mathbf{Y}$ contains the chi-square residuals $(o-e)/\sqrt{e}$ (Legendre and Legendre 1998). By applying non-centered SIMPLS to these $\mathbf{Y}$ and $\mathbf{X}$ we obtain, for axis *a*, species score vectors denoted by $\mathbf{q}_a$ and $\mathbf{r}_a$ in de Jong (1993) and ter Braak and de Jong (1998). The species scores $\mathbf{u}_{1a}$ and $\mathbf{u}_{2a}$ for axis *a* are obtained by the backtransformation

(9) $\mathbf{u}_{1a} = \mathbf{K}_1^{-1/2}\mathbf{q}_a$ and $\mathbf{u}_{2a} = \mathbf{K}_2^{-1/2}\mathbf{r}_a$.

The first axis so obtained is the same as in co-inertia analysis and maximizes Eq. (1) subject to (2). For axis $a>1$, the additional constraint is that axis *a*, derived from $\mathbf{Y}_2$, is orthogonal to the previous axes, *i.e.*

(10) $\mathbf{x}_{2a}^T\mathbf{R}_0\mathbf{x}_{2b} = 0$ for $a > b$ with $\mathbf{x}_{2a} = \mathbf{R}_2^{-1}\mathbf{Y}_2\mathbf{u}_{2a}$.

This constraint makes the difference with symmetric co-correspondence analysis.

Fitted values for the response variables are obtained in PLS by regression of $\mathbf{Y}$ on the *A* ordination axes $\mathbf{t}_1, \ldots \mathbf{t}_A$, derived from $\mathbf{X}$ (*i.e.* $\mathbf{t}_a = \mathbf{X}\mathbf{r}_a$). The values so obtained, $\hat{\mathbf{Y}}$ say, must be backtransformed to obtain fitted values for the original abundance data $\mathbf{Y}_1$. The backtransformation is

(11) $\hat{\mathbf{Y}}_1 = \mathbf{R}_1(\mathbf{R}_0^{-1/2}\hat{\mathbf{Y}}\mathbf{K}_1^{-1/2} + \mathbf{1}_n\mathbf{1}_p^T)\mathbf{K}_1$.

If for a new site *i* the abundances of predictor species are known (and the assigned weight is $\mathbf{r}_{i0}$), we can similarly obtain predicted values for the response species.

For completeness we note that symmetric co-correspondence analysis can by obtained by calculating the singular value decomposition of $\mathbf{Y}^T\mathbf{X}$, with $\mathbf{Y}$ and $\mathbf{X}$ from Eq. (8). The transformations of Eq. (9) have to be applied to the left ($\mathbf{q}_a$) and right ($\mathbf{r}_a$) singular vectors to obtain the species scores $\mathbf{u}_{1a}$ and $\mathbf{u}_{2a}$. The singular values are the square root of the eigenvalues.

Number of ordination axes and crossvalidatory fit

If $q > n$ (more predictor species than sites as is common in ecological data sets), the response data can be fitted without error by taking as many PLS-axes as there are sites, even if there is no relation between the two sets of data. Such fit has no predictive value. The number of axes to use is therefore an essential ingredient of PLS: the selected number minimizes the prediction error as estimated by cross-validation methods (Martens and Naes 1992). We applied leave-one-out crossvalidation, that is, PLS is carried out *n* times, leaving out in turn one of the sites and applying the obtained PLS model to the left-out site to predict its response species

from the predictor species. The selected number of axes is the number that minimizes the sum of squared prediction errors. We report in the example the crossvalidatory fit as 100(1- ssp_a/ssp_0) with ssp_a the sum of squared prediction errors using a PLS axes and ssp_0 the sum of squared prediction errors under the null model, in our case the row-column independence model for $\mathbf{Y}_1$. With transformation (8) and $\mathbf{R}_0 = \mathbf{R}_1$, the sums of squares of the elements of $\mathbf{Y}$ and $\mathbf{Y} - \hat{\mathbf{Y}}$ are the total inertia and the residual inertia of $\mathbf{Y}_1$, respectively. The crossvalidatory fit calculated on the transformed data is thus in terms of inertia. It can be negative when the model fits so badly that the null model predicts the data better.

There is a detail that deserves attention. When leaving out a site, the species totals and, thus, the weights $\mathbf{K}_1$ and $\mathbf{K}_2$ change (the site weights change proportionally). In our examples we use these modified weights to transform the data by Eqs (7) and (8). The left-out site is given a weight proportional to $r_{i0} = y_{1i+}$ in the complete data set. If a species occurs only once, it receives weight 0 in one of the analyses of n-1 sites. Without modification of the weights, ssp_0 is simply the total inertia of $\mathbf{Y}_1$, but with modification it is somewhat larger.

An alternative approach to select the number of axes is to test the statistical significance of each ordination axis using permutation tests. By applying the permutations to the rows of $\mathbf{Y}$ from Eq. (8), for fixed $\mathbf{X}$, we do not need to worry about differential weights, because the rows and columns of these matrices have equal weight in the analysis. The test statistic we use is the F-ratio based on the fit by the first axis to the response data (ter Braak and Šmilauer 2002). The second axis was tested by treating the first axis as covariable (*i.e.* by analyzing residuals of $\mathbf{Y}$ and $\mathbf{X}$ after fitting the first axis) so that the second axis of the original data becomes the first axis of the residualized data, and so on for further axes. Strictly speaking, this approach does not test the significance of SIMPLS axes, but of NIPALS-PLS axes (ter Braak and de Jong 1998).

In the examples we obtained the crossvalidatory fit of canonical correspondence analysis and CCA-PLS by using SIMPLS. The fit of CCA-PLS was obtained by transforming the species data $\mathbf{Y}_1$ as in Eq (8) with $\mathbf{R}_0 = \mathbf{R}_1$ and centering and standardizing (autoscaling) the environmental variables in the $\mathbf{R}_1$-metric, followed by premultiplication by the square root of $\mathbf{R}_1$ and submitting the transformed data to non-centered SIMPLS. The fit of canonical correspondence analysis was obtained by making the transformed environmental variables also orthonormal (ter Braak and de Jong 1998) by a singular value decomposition. The explained inertias of the axes so obtained were the same as those obtained with CANOCO (ter Braak and Šmilauer 2002).

Ordination diagrams

In symmetric co-correspondence analysis, ordination diagrams can be made in the usual way: by jointly plotting the species scores and site scores ($\mathbf{u}_1$ with $\mathbf{x}_1$ and $\mathbf{u}_2$ with $\mathbf{x}_2$) for the first A axes (typically $A = 2$ for easy visual inspection). As in

correspondence analysis, the interpretation of such diagrams can proceed by the centroid principle and the biplot rule (ter Braak and Verdonschot 1995, ter Braak and Šmilauer 2002). According to Eqs (1) and (2), sites are at the center of the species they contain; the diagram so conveys information on the species they are likely to contain (the centroid principle). The u-scores and x-scores together also form correspondence analysis biplots for $\mathbf{Y}_1$ and $\mathbf{Y}_2$. The interpretation of such biplots is described in full in ter Braak and Verdonschot (1995: p. 273). In brief, the correspondence analysis biplot displays for each particular species the approximate share that species has in the abundance at each site, with share defined as y_{sik}/y_{si+} (s = 1, 2 indicating the community) and, conversely, for each particular site the approximate share that site has in the abundance of each species, with share defined as y_{sik}/y_{s+k}.

Symmetric co-correspondence analysis puts emphasis on the association among species from different communities. The measure of association is the $\mathbf{R}_0$-weighted covariance between species calculated on the basis of $\mathbf{Q}$-matrices in Eq. (7); it is a covariance between relative abundances. For the optimal representation of this association in a biplot, the u-scores of each axis must be multiplied by the quarter root of the eigenvalue of the axis (this result follows from the above mentioned singular value decomposition, see also ter Braak 1990). This is important only when the multipliers differ strongly among axes. To retain the biplot interpretation for $\mathbf{Y}_1$ and $\mathbf{Y}_2$, the x-scores must be divided by the same multipliers, but we do not recommend to do so. The rescaled u-scores and original x-scores together form a Benzécri plot, an excellent compromise of conflicting aims in biplots (Gabriel 2002). In such a plot, distances among sites and among species represent chi-square distances and points of different items allow, for all practical purposes, a biplot interpretation (Gabriel 2002).

Predictive co-correspondence analysis puts emphasis on the regression of transformed $\mathbf{Y}_1$ on transformed $\mathbf{Y}_2$, and thus on the matrix of fitted values $\widehat{\mathbf{Y}}_1$ (Eq. (11)) and the rank A matrix of regression coefficients (ter Braak 1990). The fit of the response community to the predictor community ($\widehat{\mathbf{Y}}_1$) can be displayed in a correspondence analysis biplot of $\mathbf{u}_1$ and $\mathbf{x}_2$ (instead of with $\mathbf{x}_1$) and the matrix of regression coefficients by a biplot of $\mathbf{u}_1$ and $\mathbf{u}_2$ (see de Jong 1993: (37)). To infer also about the predictor community ($\mathbf{Y}_2$) one should not use $\mathbf{u}_2$ with $\mathbf{x}_2$, but the loadings of the predictor species with $\mathbf{x}_2$ (the coefficients of the $\mathbf{R}_0$-weighted regression of $\mathbf{Q}_2$ on the site scores $\mathbf{x}_2$). Together with the u-scores of the response species, the loadings of the predictor species form a biplot of their association (the covariance based on the $\mathbf{Q}$-matrices). In this biplot, distances among species represent chi-square distances. For completeness, we remark that in predictive co-correspondence analysis (and SIMPLS) the loadings of the response species (with respect to $\mathbf{x}_2$) are equal to their u-scores, and that the predictor site scores $\mathbf{x}_2$ are by default $\mathbf{R}_0$-normalized. In the examples, with A = 2 and large numbers of species, the differences between u-scores and loadings of the predictor species were not very large (r >0.97).

Since we made the distinction between u-scores and loadings, it is instructive to note why the diagrams suggested above for symmetric co-correspondence analysis ($\mathbf{u}_1$ with

$\mathbf{x}_1$ and $\mathbf{u}_2$ with $\mathbf{x}_2$) are biplots (Thioulouse et al. 1995): given the u-scores, the x-scores are optimal. For example, the scores $\mathbf{x}_1$ (Eq. (1)) are equal to the regression coefficients of the $\mathbf{K}_1$-weighted regression of $\mathbf{Q}_1^T$ on the species scores ($\mathbf{u}_1$) in the diagram, because the u-scores are orthonormal in symmetric correspondence analysis (and other forms of co-inertia analysis).

Tests on Real Data

Carabid beetles and vegetation in Dutch roadside verges

Raemakers et al. (2001) studied the relation between carabid beetles and vascular plants along roadside verges in the Netherlands. Roadside verges contribute considerably to the amount of natural area in a highly cultivated country as the Netherlands. The ultimate aim of the research is to develop management methods that increase the ecological value of roadside verges. Here we analyze the 'moist' subset of their data, comprising 30 sites with 91 carabid beetle species and 173 plant species. The beetle counts were log(y+1)-transformed. The abundances of the plant species were on a 1-9 van der Maarel scale (Jongman et al. 1995).

Both data sets are highly structured as judged from the eigenvalues and lengths of gradient (Table 2) obtained from separate ordinations by correspondence analysis (CA) and detrended correspondence analysis (DCA). Table 2 also shows the largest three eigenvalues of symmetric co-correspondence analysis (CO-CA) but these cannot be compared directly with those of CA and DCA (if $\mathbf{Y}_1 = \mathbf{Y}_2$, for example, the CO-CA eigenvalues are the square of the CA-eigenvalues; see also equation (12) in Dolédec and Chessel (1994)). The gain of directly relating beetles and plants by CO-CA over relating them through two separate analyses is expressed in terms of correlation coefficients in Table 3. The correlation between the axes of the separate analyses is high for the first axis (almost 0.9) but only moderate (<0.6) for subsequent axes (Table 3). CO-CA finds an even higher correlation on the first axis and correlations close to 0.90 for the second and third axes. High correlations on subsequent, unimportant axes may be meaningless. Therefore a further comparison is made in terms of the variance in the beetle data that is explained by ordination axes that are derived from the plant data. When such axes are obtained by CA or DCA and two of them are used for prediction, the percentage variance explained is about 15-16%, whereas the first two plant-derived axes of CO-CA explain 19% (Table 3, last column). CO-CA for relating beetles to plants thus achieves for these data a small improvement over the indirect method of relating the results of two separate analyses.

Figure 1 displays the ordination diagram of symmetric CO-CA. Selected beetle and plant species are displayed by their species scores ($\mathbf{u}_1$ and $\mathbf{u}_2$). Sites points are weighted averaged species scores. The multipliers for the u-scores to turn Figure 1 into a Benzécri plot are 0.7 and 0.6 for axes 1 and 2, respectively. For the optimal representation of beetle-plant association the aspect-ratio of Figure 1 (width : height) should thus be changed from 1 : 1 to 1: 0.85 (= 0.6/0.7 = the quarter root of the ratio of the second to the first eigenvalue). This change does not really influence the global interpretation of Figure 1: beetles and plants in corresponding positions with respect to the origin in each figure are positively associated, with stronger associations for species far from the origin. Symmetric and predictive CO-CA yield always the same first axes. In this example, the second axes are also nearly the same (r =0.98).

Raemakers et al. (2001) also classified the vegetation samples in eight syntaxonomic

units, characterized the environment by 13 variables representing soil and microclimate (acidity, moisture content in spring and summer, organic matter content, sandiness, availability of nitrate, ammonium and mineral N, soluble P and K, degree of nitrification, exposure to sun and a temperature index) and characterized the vegetation structure at each site by nine variables, including total biomass, maximum and average vegetation height, and roughness. Nearly all the quantitative variables were log-transformed to make their distributions less skew. Separate canonical correspondence analyses (CCA) of the beetle data with respect to these three data sets explained, in two dimensions, 19%, 18% and 13% of the beetle variance, which can be compared with the 19% explained by CO-CA. In four dimensions these figures are 27, 28, 22 and 28%, respectively, and in the maximum dimension in each analysis 33, 51, 34 and 100%. Clearly, comparing these statistics has little meaning because the numbers of predictor variables differ. To place the analyses on equal footing we applied leave-one-out crossvalidation to determine the optimal dimension for each (Figure 2). In the crossvalidation we used predictive CO-CA. All percentages are low, but that is inherent to data that are largely qualitative (presence versus absence), rather than quantitative. CO-CA has a local maximum (7.8%) for two axes and a global maximum (8.7%) at seven axes. We retain the two-dimensional solution to keep the model as simple as possible. When using CO-CA, the plants thus predict 7.8% of the beetle inertia. With the optimal dimension in brackets, the syntaxa predict 8.1% using CCA (2), the environmental variables predict 4.8% using CCA (2) and 3.1% using CCA-PLS (1), whereas vegetation structure gives negative percentages and thus has no predictive value for the beetle community using either CCA or CCA-PLS. In conclusion, the vegetation either expressed as abundances of individual species or as syntaxonomic units, predicted the beetle data better than the environmental measurements, and vegetation structure had no predictive value for the beetle data in this study.

The first two axes of CO-CA were significant ($P<0.001$), whereas subsequent axes were not ($P>0.60$). In Figure 1 the horizontal axis is positively correlated with the moisture content variables and organic matter (all $r \approx 0.6$), the vertical is negatively correlated with acidity ($r = -0.43$). The CCAs with respect to the syntaxonomic data and the environmental data also showed two significant axes each ($P=0.002$ and 0.01 for the syntaxonomic data and $P=0.001$ and 0.038 for the environmental data), whereas further axes were not ($P>0.08$). With the structure variables none of the CCA axes was significant ($P>0.30$). The number of significant axes thus agreed in this study with the optimal dimension chosen by leave-one-out crossvalidation.

The use of individual plant species to predict the beetle community avoids the, sometimes debatable, classification of sites into syntaxonomic units. But, because the analyses showed that there is not much to be gained by using all plant species to predict the beetle community, we may equally well conclude that it is sufficient in this case to summarize the plant species composition into syntaxomonic units, and also that it is unlikely that a modification of the vegetation classification would result in a better fit to the beetle community.

Bryophytes and vascular plants in Carpathian spring meadows

Hájek, Hekera and Hájková (2002) explored data on the community composition of bryophytes and vascular plants in 70 spring meadows in the Western Carpathians and environmental correlates expressed as 14 water-chemical variables and one variable represented slope. We log-transformed all concentration variables. The data are also analyzed in case study 2 of Lepš and Šmilauer (2003). For practical purposes we limited the data to species that occur 5 or more times, giving 30 bryophyte species and 123 vascular plants. The data contain a very strong first gradient as judged by DCA (5.2 SD for the bryophytes and 3.0 SD for the vascular plants), whereas the second gradient is ~2.5 SD for both data sets. The first axes of the two analyses are strongly correlated ($r \approx 0.9$) whereas the second axes are not ($r \approx 0.1$) (Hájek et al. 2002). Correspondence analyses show clearly arched configurations of species and sites points for these data. Figure 3, based on predictive CO-CA of the bryophytes against the vascular plants, illustrates that CO-CA also suffers from the arch effect. We have not yet implemented a detrended version.

For predictive purposes, detrending may not be necessary. Figure 4 shows that the bryophytes are better predicted by the vascular plants (28% with 5 CO-CA axes) than by the environmental variables (17% with 2 CCA axes and 18% with 4 CCA-PLS axes). In terms of crossvalidatory fit, the arched two-dimensional CO-CA (Figure 3) accounts already for 25% of the bryophyte inertia. Only these two axes are significant ($P = 0.001$), whereas subsequent axes are not ($P > 0.10$). In the CCA the first two axes are significant (P=0.001 and 0.04), whereas the third is not (P =0.52).

Discussion

No direct quantitative method existed so far for predicting one biological community from another. Co-correspondence analysis fills this gap. It fits in the weighted averaging family of methods for ecological gradient analysis (Table 1; ter Braak and Prentice 1988), of which weighted averaging, (detrended) correspondence analysis and canonical correspondence analysis are the most widely applied (Birks et al. 1996). Within this family it has a well-defined domain of application; it is the direct ordination method for relating one community data set to another (Table 1). It improves upon the indirect method of correlating the ordination axes of separate (detrended) correspondence analyses of the data sets, the main improvement being that co-correspondence analysis can be used in a predictive way, as our examples show.

Co-correspondence analysis combines the ecological method of weighted averaging and the chemometric method of partial least squares, and derives data-analytic strength from both. By maximizing the covariance between the weighted averaged species scores of one community with those of the other community, co-correspondence analysis attempts to identify the ecological gradients that are common to both communities. To further clarify the logic of the method, let us assume for a moment that these ecological gradients coincide with particular environmental

variables. If measured, these could be used to predict one community (or both) through canonical correspondence analysis. But, what if in a future application the environmental measurements were missing and we want to infer about the first community from the second one? The data on the second community could then be used to infer about the environmental data – a calibration problem (ter Braak and Prentice 1988) - and the inferred environmental data could then be entered to canonical correspondence analysis to predict the first community. Herein, the calibration problem could be solved with WA-PLS (Table 1). This method of predicting one community from another thus consists of two steps. Clearly, co-correspondence analysis is a one-step method: it integrates the calibration method of WA-PLS with the constrained ordination method of canonical correspondence analysis. The second (predictor) community is used as a multivariate bio-assay for the true underlying gradients. In practice, the environmental basis of the ecological gradients is not precisely known – that is why environmental measurements may yield worse predictions than the second community's data, as in our examples.

Here we resolved the problem of differences in implied site weights (see Introduction) by working from first principles: the maximization of the weighted covariance between weighted averaged species scores. This resolved the weight problem in a versatile way. Rather than being implied, the weights for use in the covariance (denoted by $\mathbf{R}_0$) can be chosen freely by the user. As the choice influences the results of the analysis, this brings up a new problem: which weights to choose? In the examples we chose to use the implied weights of the community in the response role ($\mathbf{R}_0 = \mathbf{R}_1$) in order to facilitate the comparison with canonical correspondence analysis. This is also the logical choice in regression (Seber 1977) - and thus in predictive co-correspondence analysis and canonical correspondence analysis - when these site weights are indicative for the precision of the abundance data of the response species. This is typically the case for count data and presence/absence (1/0) data with low incidence probability (Jongman et al. 1995). With this choice, the scores of the response species are (proportional to) weighted averages of site scores derived from the predictors (Eq. (3)), precisely as in canonical correspondence analysis (ter Braak 1986). This choice thus makes co-correspondence analysis closest to a (canonical) correspondence analysis of the response species. In symmetric co-correspondence analysis, a symmetric choice has perhaps more appeal, *e.g.* $\mathbf{R}_0 = (\mathbf{R}_1+\mathbf{R}_2)/2$. We propose the term co-correspondence analysis to default to predictive co-correspondence analysis using SIMPLS and $\mathbf{R}_0 = \mathbf{R}_1$.

Co-correspondence analysis inherits not only the good properties from correspondence analysis, but also the bad ones, most notably the arch effect (Jongman et al. 1995, Legendre and Legendre 1998). The arch effect – the effect that the second axis' scores show a systematic, often parabolic, relation with the first axis' scores (Figure 3) - may hamper ecological understanding of the underlying gradients. The proposed remedy, detrending (Hill and Gauch 1980), can in principle be applied in co-correspondence analysis, but we have not yet attempted to apply it. Detrending is no panacea - it may flatten out some real variation (Minchin 1987) - and is therefore rather controversial (Legendre and Legendre 1998). Moreover, detrending is not

necessary for prediction purposes as we showed in our second example.

The transformation in Eq. (8) is akin to one of the ecologically meaningful transformations for ordination of species data proposed by Legendre and Gallagher (2001). Their idea was to transform the community data in such a way that the Euclidean distance between sites after transformation is identical to the chi-square distance between the sites before transformation. The transformation that achieves this is $y^*_{2ik} = y_{2ik} / (y_{2i+}\sqrt{y_{2+k}})$. Pinel-Alloul et al. (1995) applied this transformation to phytoplankton and fish data that were used as predictors in a canonical correspondence analysis of a zooplankton community. This analysis implies row weights $\{y_{1i+}\}$. The resulting transformation is almost identical to **X** (and also **Y**) in Eq. (8) with $\mathbf{R}_0 = \mathbf{R}_1$ (there is a small difference in the implied centering of $\mathbf{Y}_2$). What makes their method really different from ours is that Pinel-Alloul et al. (1995) used forward selection (instead of PLS) to overcome the problem that the number of predictor species was greater than the number of sites. Selection of predictor species has four disadvantages: (1) it complicates the formal statistical testing of the association between the communities, (2) it does not aim to preserve (chi-square) distances among sites in the predictor space, (3) it destroys the logic of the original transformation because the total abundance of the selected species is not equal to that of all predictors species, and (4) it does not use all the information available in the predictor set. In particular, their method does not have the weighted averaging properties (3) – (6) of our method.

Co-correspondence analysis may help in the search for good indicators for biodiversity (Noss 1990). Not all species groups are equally easy to sample or identify. One could try to predict a difficult species group from an easy one. The building of the prediction model with co-correspondence analysis requires representative training data for both species groups from a common set of sites, but from there only the easy group need to be sampled and identified. Co-correspondence analysis may help in the search for the most suitable indicators.

Acknowledgements

We thank Dean Fairbanks and Karlè Sýkora for their questions that led to this research, Ivo Raemakers and Michal Hájek for permission to use their data and Petr Šmilauer, Hilko van der Voet, Theo Reijmers, John Birks, Richard Furnas, Pierre Legendre and the subject editor Norman Kenkel for comments on the manucript.

Literature cited

Birks, H. J. B. 1998. Numerical tools in palaeolimnology - progress, potentialities, and problems. Journal of Paleolimnology 20:307-332.

Birks, H. J. B., S. M. Peglar, and H. A. Austin. 1996. An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986-1993. Abstracta Botanica 29:17-36.

Birks, H. J. B., J. M. Line, S. Juggins, A. C. Stevenson, and C. J. F. ter Braak. 1990. Diatoms and pH reconstruction. Philosophical Transactions of the Royal Society London, Series B **327**:263-278.

Burnham, A. J., R. Viveros, and J. F. MacGregor. 1996. Frameworks for latent variable multivariate regression. Journal of Chemometrics **10**:31-45.

de Jong, S. 1993. SIMPLS: an alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18:251-263.

de Jong, S., and C. J. F. ter Braak. 1994. Comments on the PLS kernel algorithm. Journal of Chemometrics **8**:169-174.

Dolédec, S., and D. Chessel. 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology 31:277-294.

Frisvad, J. C., and M. Norsker. 1996. Use of correspondence analysis partial least squares on linear and unimodal data. Journal of Chemometrics 10:677-685.

Fritz, S. C., S. Juggins, R. W. Battarbee, and D. R. Engstrom. 1991. Reconstruction of past changes in salinity and climate using a diatom-based transfer function. Nature **352**:706-708.

Gabriel, K. R. 2002. Goodness of fit of biplots and correspondence analysis. Biometrika 89:423-436.

Hájek, M., P. Hekera, and P. Hájková. 2002. Spring fen vegetation and water chemistry in the Western Carpathian flysch zone. Folia geobotanica 37:205-224.

Hill, M. O. 1973. Reciprocal averaging: an eigenvector method of ordination. Journal of Ecology 61:237-249.

Hill, M. O., and H. G. Gauch. 1980. Detrended correspondence analysis, an improved ordination technique. Vegetatio 42:47-58.

Jongman, R. H. G., C. J. F. ter Braak, and O. F. R. van Tongeren. 1995. Data analysis in community and landscape ecology. Cambridge University Press, Cambridge.

Legendre, L., and P. Legendre. 1998. Numerical ecology. Elsevier, Amsterdam.

Legendre, P., and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia **129**:271-280.

Lepš, J., and P. Šmilauer. 2003. Multivariate Analysis of Ecological Data using CANOCO. Cambridge University Press.

Magnus, J. R., and H. Neudecker. 1988. Matrix differential calculus with applications in statistics and econometrics. Wiley, New York.

Martens, H. 1986. "Unscrambler" program for calibration. CAMO, Trondheim, Norway (www.camo.no).

Martens, H., and T. Naes. 1992. Multivariate calibration. Wiley, Chicester.

MATLAB. 2000. Using MATLAB version 6. The Math Works, Inc., Natrick, MA, USA (www.mathworks.com).

Minchin, P. R. 1987. An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 67:1167-1179.

Noss, R. F. 1990. Indicators for monitoring biodiversity: a hierarchical approach. Conservation Biology 4:355-364.

Persson, S. 1981. Ecological indicator values as an aid in the interpretation of ordination diagrams. Journal of Ecology **69**:71-84.

Pinel-Alloul, B., T. Niyonsenga, and P. Legendre. 1995. Spatial and Environmental Components of Fresh-Water Zooplankton Structure. Ecoscience **2**:1-19.

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

Raemakers, I. P., A. P. Schaffers, K. V. Sýkora, and T. Heijerman. 2001. The importance of plant communities in road verges as habitat for insects. Proceedings of the Section Exprimental and Applied Entomology of the Netherlands Entomological Society 12:101-106.

Seber, G. A. F. 1977. Linear regression analysis. Wiley, New York.

ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167-1179.

ter Braak, C. J. F. 1990. Interpreting canonical correlation analysis through biplots of structural correlations and weights. Psychometrika 55:519-531.

ter Braak, C. J. F., and S. de Jong. 1998. The objective function of partial least squares regression. Journal of Chemometrics 12:41-54.

ter Braak, C. J. F., and S. Juggins. 1993. Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. Hydrobiologia 269: 485-502.

ter Braak, C. J. F., S. Juggins, H. J. B. Birks, and H. Van der Voet. 1993. Weighted averaging partial least squares regression (WA-PLS): definition and comparison with other methods for species-environment calibration. Pages 525-560 in G. P. Patil and C. R. Rao, editors. Multivariate Environmental Statistics. North-Holland, Amsterdam.

ter Braak, C. J. F., and P. Šmilauer. 2002. CANOCO Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination (version 4.5). Microcomputer Power, Ithaca, NY, USA (www.canoco.com).

ter Braak, C. J. F., and I. C. Prentice. 1988. A theory of gradient analysis. Advances in Ecological Research 18:271-317.

ter Braak, C. J. F., and H. van Dam. 1989. Inferring pH from diatoms: a comparison of old and new calibration methods. Hydrobiologia **178**:209-223.

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.

Thioulouse, J., D. Chessel, and S. Champely. 1995. Multivariate analysis of spatial patterns: a unified approach to local and global structures. Environmental and Ecological Statistics **2**:1-14.

Thioulouse, J., S. Dolédec, D. Chessel, and J. M. Olivier. 1995. ADE-4 Program Library. Multivariate analysis and graphical display for environmental data. http://pbil.univ-lyon1.fr/ADE-4/.

Tucker, L. R. 1958. An inter-battery method of factor analysis. Psychometrika 23:111-136.

Wise, B. M., and N. B. Gallagher. 2000. PLS_Toolbox 2.1. Eigenvector Research, Inc, Manson, WA, USA (www.eigenvector.com).

Wold, S. 2002. SIMCA version 10. Umetrics, Kinnelon, NJ USA (www. umetrics. com).

Appendix A. How to carry out symmetric co-correspondence analysis in ADE-4 (Thioulouse et al. 1995).

1. Divide the data matrices by their overall totals, yielding new data matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Calculate and save the row totals ($\mathbf{R}_1$ and $\mathbf{R}_2$) and column totals ($\mathbf{K}_1$ and $\mathbf{K}_2$) of $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

2. Decide on $\mathbf{R}_0$, for example, $\mathbf{R}_0 = (\mathbf{R}_1+\mathbf{R}_2)/2$ or $\mathbf{R}_0 = \mathbf{R}_1$.

3. Calculate the chi-square residuals as indicated in and below Eq. (7), either in a spreadsheet or in ADE-4, yielding the data matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$.

4. Carry out two non-centered principal component analyses via the 'PCA' section, one on $\mathbf{Q}_1$ using row weights $\mathbf{R}_0$ and column weights $\mathbf{K}_1$ and one on $\mathbf{Q}_2$ using row weights $\mathbf{R}_0$ and column weights $\mathbf{K}_2$. In each analysis, save all axes with non-zero eigenvalues.

5. Carry out the co-inertia analysis via the 'CoInertia' section. From this menu, first choose 'Matching two statistical triplets' and specify as input files the two output files with extension '.ncta' of step 3. This yields an output file with extension '.iita'. Secondly, choose 'Coinertia analysis' with as input file the iita-file to run the actual analysis.

Table 1. Overview of gradient analysis methods based on weighted averaging.

The community data consist of incidences or abundances (≥0) of a set of species at a set of sites. The environmental variables, measured at the same set of sites, are quantitative and/or qualitative (0/1). The methods use weighted averages of species scores, appropriate for unimodal data, and linear combinations of environmental variables, appropriate for linear data (ter Braak and Prentice 1988).

Method	Abbreviation	Response variables	Predictors
Correspondence analysis	CA	Community data	-
Canonical correspondence analysis	CCA	Community data	Environmental variables
CCA partial least squares	CCA-PLS	Community data	Many environment variables
Weighted averaging calibration	WA	Environmental variable	Community data
WA partial least squares	WA-PLS	Environmental variable(s)	Community data
Co-correspondence analysis	CO-CA	Community data	Community data

Table 2. Eigenvalues of the first three axes of separate CAs and DCAs and of symmetric CO-CA of beetles and plants. The total inertia (sum of all eigenvalues of CA) is 4.99 for beetles and 5.65 for plants. The sum of all eigenvalues of CO-CA is 0.94.

			Axis	
	Method	1	2	3
Beetles	CA	0.50	0.36	0.32
	DCA	0.50	0.32	0.21
	Length of gradient	3.22	2.74	2.57
Plants	CA	0.57	0.53	0.42
	DCA	0.57	0.41	0.27
	Length of gradient	3.44	2.99	2.88
Beetles-plants	CO-CA	0.25	0.13	0.08

Table 3. Correlation coefficients between beetle-derived and plant-derived site scores of the first three axes of separate CAs and DCAs and of symmetric CO-CA (%fit = the percentage fit of the beetle data by the first two plant-derived axes). Site weights are beetle-based.

Method		Axis		%fit
	1	2	3	
CA	0.88	0.27	0.46	15
DCA	0.89	0.53	0.07	16
CO-CA	0.96	0.94	0.88	19

Legends to Figures

Figure 1. Biplot based on symmetric co-correspondence analysis of carabid beetles (left: $\mathbf{u}_1$ and $\mathbf{x}_1$) and plants (right: $\mathbf{u}_2$ and $\mathbf{x}_2$) in roadside verges showing ~19% of the total variance of each data set. The species displayed (triangle) have a more than average fit and occur five or more times in the data. Symbols of sites indicate their syntaxonomic unit. Scores of species and sites are scaled according to Eqs (1) and (2). The syntaxonomic units are (approx. from left to right) Molinio-Arrhenatheretea / Koelerio-Corynephoretea (black rectangle), Tanaceto-Artemisietum / Arrhenatheretalia (squares), Arrhenatheretum medicaginetosum (gray circles), Arrhenatheretum elatiorus (stars), Galio-Alliarion / Arrhenatherion (gray rectangle), Galio-Alliarion / Alopecurion (diamond), Valeriano-Filipenduletum (open circle), Molinietalia (Calthion)(down triangle). The carabid beetles shown are: ACUPPARV = Acupalpus parvulus, AMARAAEN = Amara aenea, AMARABIF = Amara bifida, AMARALUN = Amara lunicollis, ANISOBIN = Anisodactylus binotatus, BEMBIPRO = Bembidion properans, BRADYHAR = Bradycellus harpalinus, CALATFUS = Calathus fuscipes, CALATMEL = Calathus melanocephalus, CARABGRA = Carabus granulatus, CARABMON = Carabus monilis, DYSCHGLO = Dyschirius globosus, HARPAAFF = Harpalus affinis, HARPARUF = Harpalus rufibarbis, LORICPIL = Loricera pilicornis, PTEROANT = Pterostichus anthracinus, PTEROMEL = Pterostichus melanarius, PTEROMIN = Pterostichus minor, PTERONIG = Pterostichus niger, PTERONIH = Pterostichus nigrita, PTEROSTR = Pterostichus strenuus, TRECH-SP = Trechus species. The plant species shown are: Achimill = Achillea millefolium, agrocapi = Agrostis capillaris, agrostol = Agrostis stolonifera, alopprat = Alopecurus pratensis, bellpere = Bellis perennis, bracruta = Brachythecium rutabulum, callicus = Calliergonella cuspidata, cardprat = Cardamine pratensis, cerafont = Cerastium fontanum, festrubr = Festuca rubra, glechede = Glechoma hederacea, heraspho = Heracleum sphondylium, holclana = Holcus lanatus, lotucorn = Lotus corniculatus, phalarun = Phalaris arundinacea, phraaust = Phragmites australis, planlanc = Plantago lanceolata, poa prat = Poa pratensis, poa triv = Poa trivialis, ranubulb = Ranunculus bulbosus, rhytsqua = Rhytidiadelphus squarrosus, senejaco = Senecio jacobea, tanavulg = Tanacetum vulgare, trifdubi = Trifolium dubium.

Figure 2. Crossvalidatory fit of the beetle data set against the number of ordination axes for different predictor data sets (plants: closed squares; syntaxa: open squares; environment: triangles; vegetation structure: circles) and methods (predictive CO-CA: solid; CCA: long dash; CCA-PLS: short dash).

Figure 3. Biplot based on predictive co-correspondence analysis of bryophytes (left) against vascular plants (right) in Carpathian spring meadows, showing 30% and 21% of the total variance in the bryophyte data and vascular plant data, respectively. The species (triangles) are positioned in the graph according to their loadings with respect to normalized site scores (circles) derived from the vascular plants ($\mathbf{x}_2$). Only vascular plants with more than average fit are shown. Full names of bryophytes and vascular plants are given in Hájek et al. (2002).

Figure 4. Crossvalidatory fit of the bryophyte data set against the number of ordination axes for different predictor data sets (higher plants: squares; environment: triangles) and methods (predictive CO-CA: solid; CCA: long dash; CCA-PLS: short dash).

Figure 1. Biplot based on symmetric co-correspondence analysis of carabid beetles (left:u_1 and x_1) and plants (right: u_2 and x_2) in roadside verges showing ~19% of the total variance of each data set. The species displayed (triangle) have a more than average fit and occur five or more times in the data. Symbols of sites indicate their syntaxonomic unit. Scores of species and sites are scaled according to Eqs (1) and (2).

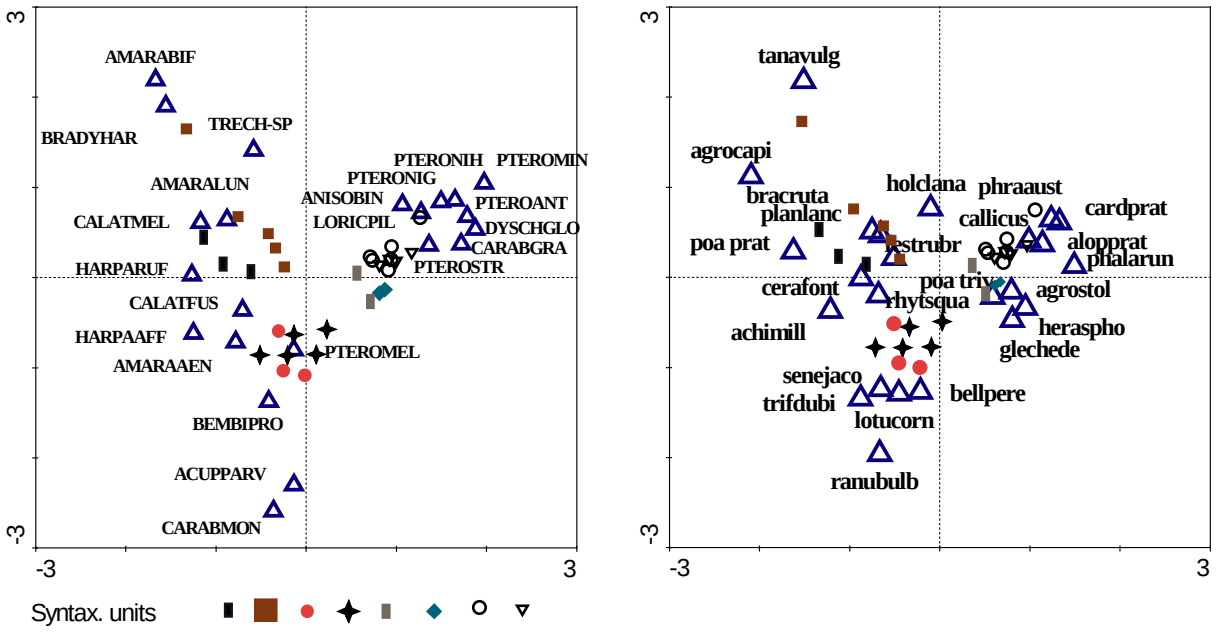

Figure 2. Crossvalidatory fit of the beetle data set against the number of ordination axes for different predictor data sets (plants: closed squares; syntaxa: open squares; environment: triangles; vegetation structure: circles) and methods (predictive CO-CA: solid; CCA: long dash; CCA-PLS: short dash).

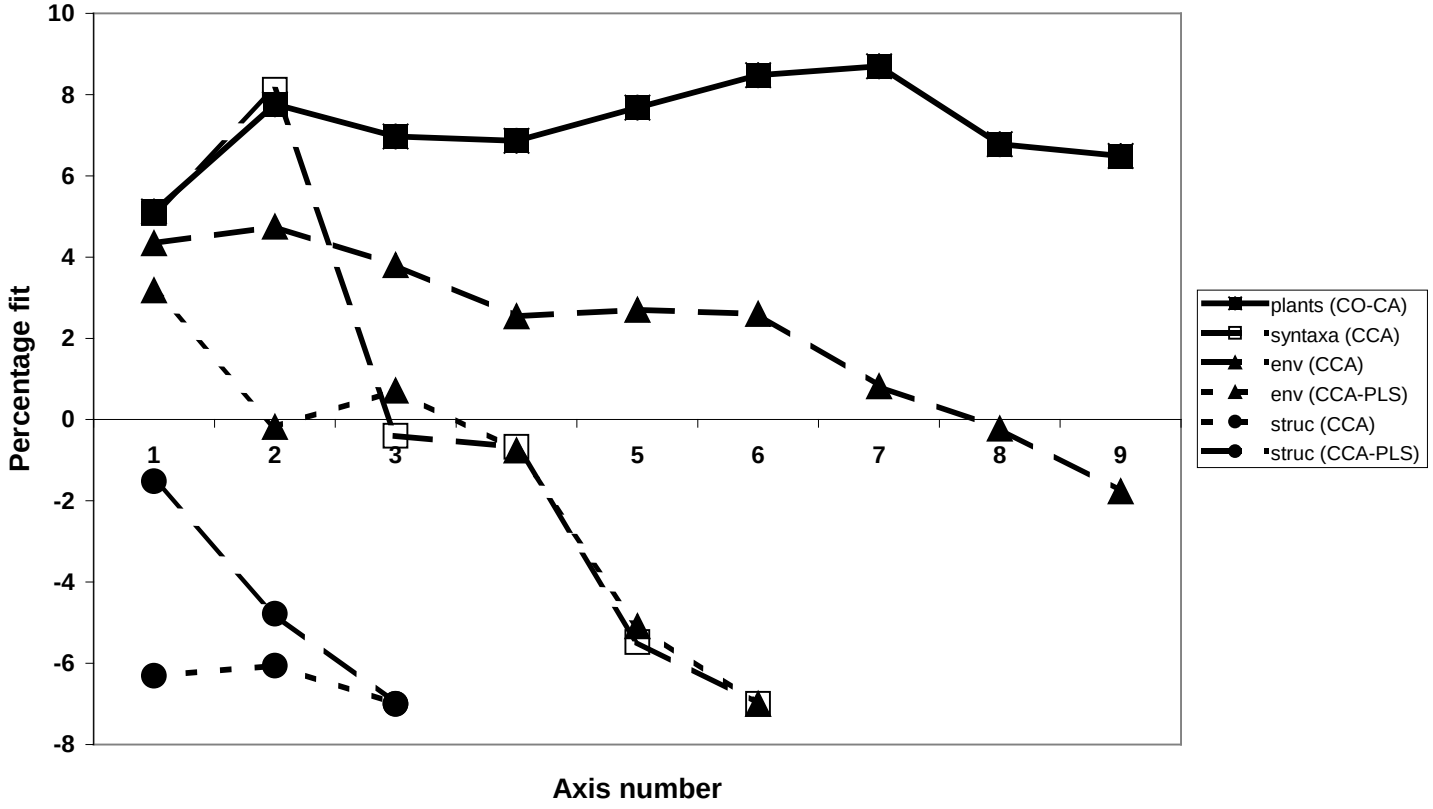

Figure 3. Biplot based on predictive co-correspondence analysis of bryophytes (left) against vascular plants (right) in Carpathian spring meadows, showing 30% and 21% of the total variance in the bryophyte data and vascular plant data, respectively. The species (triangles) are positioned in the graph according to their loadings with respect to normalized site scores (circles) derived from the vascular plants (x_2). Only vascular plants with more than average fit are shown. Full names of bryophytes and vascular plants are given in Hájek et al. (2002).

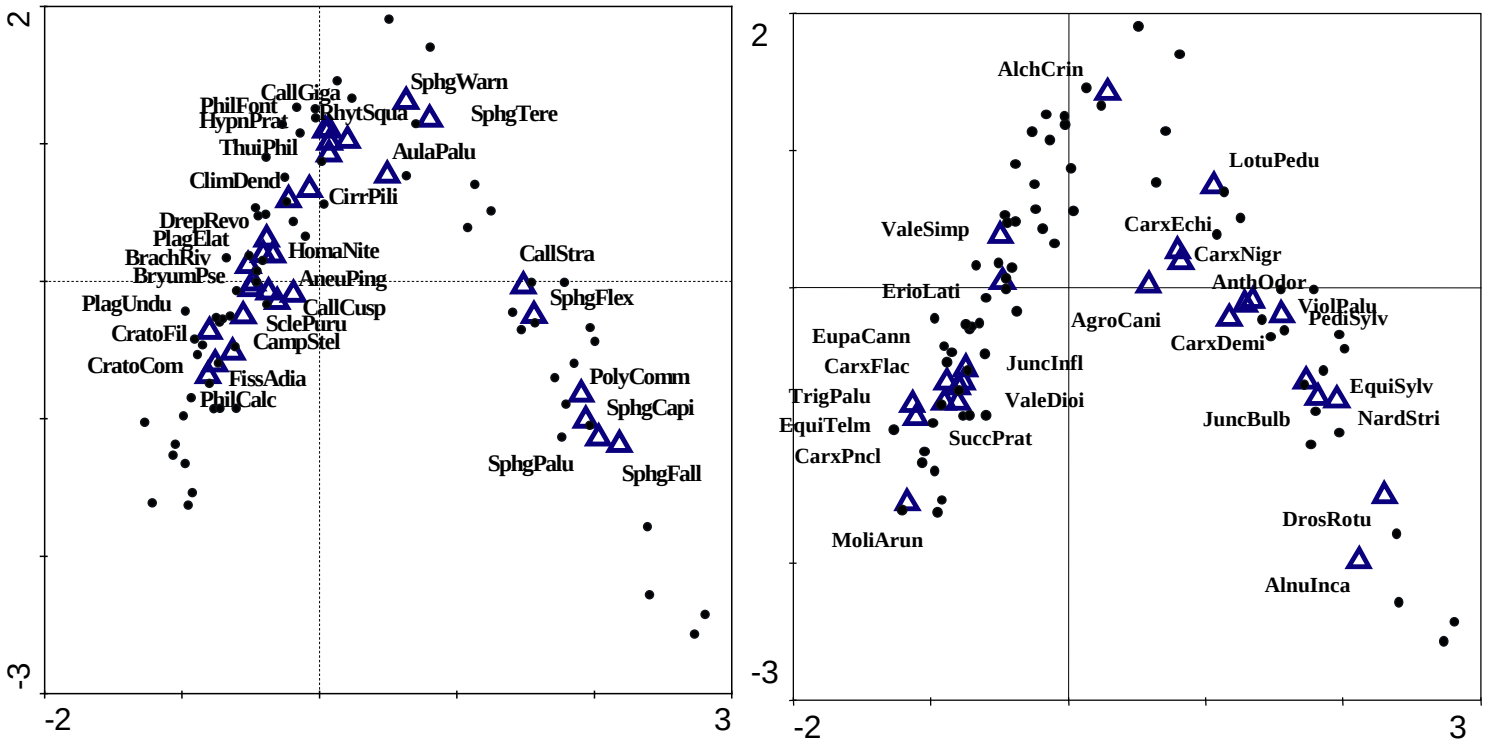

Figure 4. Crossvalidatory fit of the bryophyte data set against the number of ordination axes for different predictor data sets (higher plants: squares; environment: triangles) and methods (predictive CO-CA: solid; CCA: long dash; CCA-PLS: short dash).