Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Chapter 15 Offprint

Canonical Correspondence Analysis

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación BBVA

CHAPTER 15

Canonical Correspondence Analysis

PCA, CA and LRA operate on a single data matrix, and have similar ways of reducing the high dimensionality of the data to a low-dimensional approximation for ease of interpretation. The low-dimensional views of the data are the best in terms of the least-squares criterion in each case, accounting for a maximum amount of variance while simultaneously minimizing the unexplained variance. Often additional data are available, which can be related afterwards to an existing ordination. One of the most common situations in ecology is when the data consist of biological measurements (e.g., species abundances) at different locations, and in addition there are various environmental variables observed at the locations. We have shown how biological data can be optimally displayed with respect to ordination axes and then how the environmental variables can be related to these dimensions. The reverse can also be done, first optimally displaying the environmental data and then fitting the biological data to the results. In either case these relationships might be weak. Ecologists may be more interested in that part of the biological information that is more directly related to the environmental information. This environmentally related part of the biological variance is also multidimensional, so we again resort to ordination to interpret it through dimension reduction. Methods that relate two sets of data are often described as *canonical* in statistics, and this chapter deals mainly with one of the most popular in ecology, canonical correspondence analysis.

Contents

Response and explanatory variables 190
Indirect gradient analysis 190
Direct gradient analysis 192
Restricted ordination in PCA and LRA 194
CCA as the analysis of weighted averages 194
Coding of explanatory variables 195
CCA as a discriminant analysis 197
SUMMARY: Canonical correspondence analysis 199

Response and explanatory variables

In Chapter 10 we looked at the introductory data set "bioenv" in detail and made regression biplots using two of the environmental variables, pollution and depth, as the support of the biplot, or two so-called *canonical dimensions* that were obtained by maximizing the correlation between the biological and environmental data sets. Since then we have learned a bit more about analysing abundance data using CA, so in this chapter we will introduce a variant of CA, called *canonical correspondence analysis* (CCA), which is appropriate for this particular combination of biological and environmental measurements on the same samples. Recalling Exhibit 1.1, there were 30 samples and the five biological variables, regarded as response variables, accompanied by four environmental variables, of which three are on continuous scales and one on a categorical scale. The objective is to find out how much of the variance (in the CA sense, in other words, inertia) is accounted for by the environmental variables and to interpret the relationship. In this approach the two sets of variables are considered asymmetrically: the biological data are the responses (like the *"Y"* variables in a regression) and the environmental variables are the explanatory variables, or predictors (like the *"X"* variables). This is different from the canonical correlation analysis of Chapter 10, which treated the two sets of data symmetrically and would have been the same if the two sets of variables were interchanged.

Indirect gradient analysis

Before explaining CCA, let us first consider the CA of the 30×5 matrix of biological data and, as before, the ways of displaying the environmental variables on the CA biplot. Exhibit 15.1 shows the row-principal contribution biplot of the samples and species, which means that the distances between the samples are approximate chi-square distances between their profiles, and the standardized species (standardized in the CA sense – see Chapter 13) have been regressed by weighted least-squares on the two CA dimensions and are depicted by their regression coefficients. The total inertia of the biological data is equal to 0.544 and the axes account for 0.288 (53.0%) and 0.121 (22.2%) respectively, that is 75.2% of the total. Both 75.2% of the variance of the sample points and 75.2% of the variance of the species is explained by this solution.

While the variance explained for the species abundance data is the best possible according to the optimization criterion in CA, the regressions of the environmental variables are much lower and can vary a lot in terms of variance explained:

Depth	30.4%
Pollution	69.5%
Temperature	2.1%
C (clay)	3.4%
S (sand)	9.9%
G (gravel)	18.7%

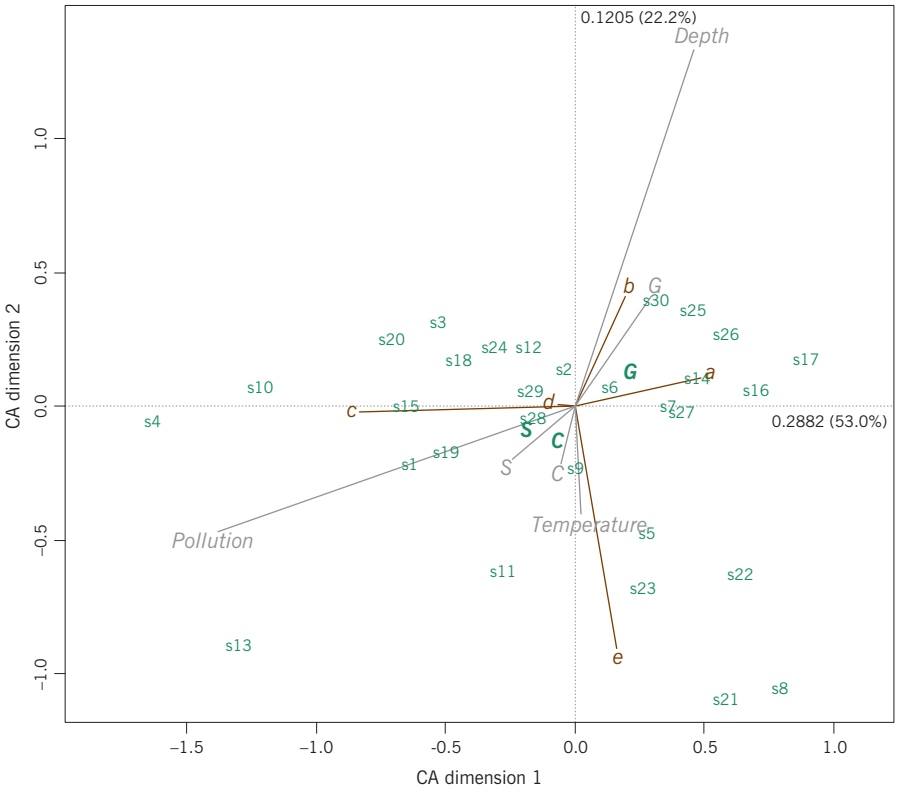


Exhibit 15.1: *CA biplot of the biological data in the "bioenv" data set, with samples in principal coordinates and species in contribution coordinates. The one discrete and three continuous environmental variables are shown according to their regression coefficients and the discrete variable's categories are additionally shown (in black) at the centroids of the samples in the corresponding categories*

Notice that dummy variables such as the sediment categories *C*, *S* and *G* (the gray points in Exhibit 15.1), with values of 0 and 1, will always have low variance explained. The other way of showing the categories is as centroids of the samples (the black points in Exhibit 15.1) – this can be achieved by adding three extra rows to the data matrix where the abundances of the species are aggregated across the samples for each sediment type, and declaring these additional rows as *supplementary points*. These additional rows are as follows:

	a	*b*	*c*	*d*	*e*
C	105	46	73	81	27
S	103	70	115	104	32
G	196	146	64	142	30

The (row) profiles of the sediment categories are exactly the centroids shown in Exhibit 15.1. These centroids do not lie on the same vector as the dummy vari-

ables, but there is a close mathematical relationship between these alternative sets of coordinates for category points added to the display, which depends on the mass of each category and the parts of inertia on each axis. In any case, what we are assured of is that the corresponding categories always lie in the same quadrant (one of the four regions defined by the two ordination axes), and if the parts of inertia are similar, then the centroids will lie very close to the dummy variable biplot axis.

This type of analysis in Exhibit 15.1 is called *indirect gradient analysis*: first an ordination is obtained optimally displaying the samples and response variables (here, the species), and then the explanatory variables are related to the ordination axes.

Direct gradient analysis

In indirect gradient analysis the relationship between the explanatory variables and the response variables is conditioned on the ordination of the response variables. One could imagine a situation in which the main dimensions of the responses have little relationship with the explanatory variables, because this relationship is to be found on less important dimensions of the response data. So, in order to focus specifically on the relationship between biological and environmental variables in this example, we first make a projection of the biological variables into the space of the environmental variables. This is also called *constrained* or *restricted* ordination, because a condition is introduced that the ordination axes must be linear functions of the environmental variables.

There are three continuous variables and three dummy variables, but – as in regression analysis – the dummy variables count for one less because of their interdependency, so there are five dimensions in the explanatory variable space. The first step then is to project the species response data into this space, which also means we eliminate all variance in the response data that is not correlated linearly with the explanatory variables – we are only interested in that part of the variance that is correlated with the environmental variables. The total inertia of the species data was, as we reported earlier, 0.544, and it turns out that the amount 0.249 of this inertia is linearly related to the environmental variables, i.e. 45.8% of the total. So from now on we are only interested in this constrained part of the inertia.

The analysis then continues as a regular CA in this restricted five-dimensional space, to find the axes that explain a maximum of this constrained inertia – Exhibit 15.2 shows the result of what is now a *canonical correspondence analysis* (CCA), in the form of a *triplot* of samples, species and environmental variables. Almost all (96.3%) of this constrained inertia of 0.249 is explained in the new ordination map. Pollution is the most important variable on the first axis,

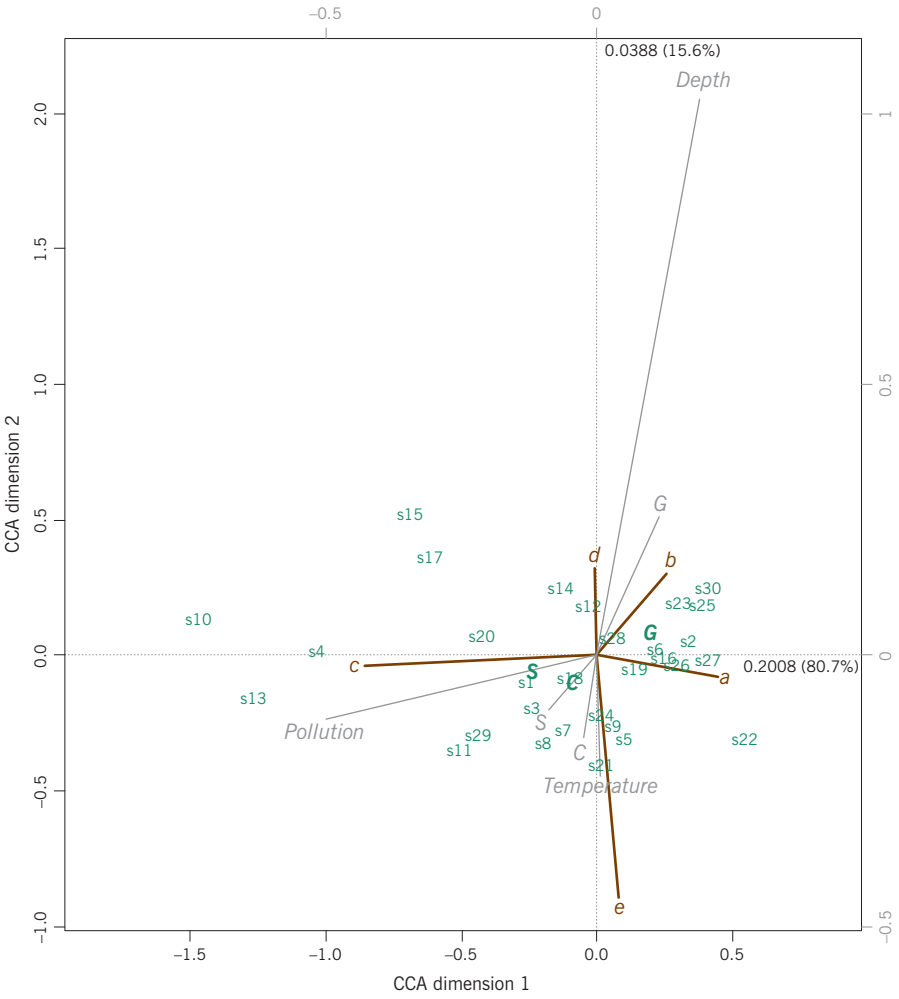


Exhibit 15.2:
Canonical correspondence analysis triplot of "bioenv" data. The row-principal scaling with species in contribution coordinates is again shown, as well as the environmental variables regressed onto the ordination axes. Percentages of inertia explained are with respect to the restricted inertia

whereas depth is the most important on the second. Because the regressions of the environmental variables are performed on the sample principal coordinates that have much less variance on the second axis, depth's regression coefficient on the second axis is large and the variable gives the impression that it is more important than pollution. If we wanted comparability between the coordinates of the gradient vectors of the environmental variables, the biplot should be made using the standard coordinates of the samples, as in the regression biplots of Chapter 10.

The variances explained by the CCA axes of the environmental variables are now much higher than before, due to the constraining of the solution:

Depth	85.3%
Pollution	99.1%
Temperature	3.3%
C (clay)	8.8%
S (sand)	14.8%
G (gravel)	33.5%

Notice again that the dummy variables, by their very nature of having only values of 0 and 1, cannot attain a high percentage of variance explained – this issue is dealt with in the next chapter.

In the CCA described above, we have analysed the inertia of the abundance data in the space constrained by the environmental variables. The constrained space is formed by axes that are predictions based on linear regression on the environmental variables. In other applications researchers might be more interested in the inertia not correlated with a particular environmental variable or variables – for example, it may be of interest to partial out a known effect such as a latitudinal gradient. Exploring this unconstrained part is called *partial CCA*, which will be illustrated in the case study of Chapter 19.

Restricted ordination in PCA and LRA

The constrained version of CA illustrated above is similarly applicable to PCA and LRA. When the responses are continuous variables on an interval scale, then the version of PCA restricted in terms of a separate set of explanatory variables is called *redundancy analysis*. Similarly, when the responses are compositional data and LRA is applicable, it is possible to restrict the solution to be linearly related to predictor variables. The idea is the same in each case: project the data, with its particular distance function, into the space of the explanatory variables, and then carry on as before. We continue with CCA, which is the most popular of these options.

CCA as the analysis of weighted averages

There is another way of thinking about CCA, in terms of the weighted averages of the explanatory variables, using the relative abundances of the species as weights. Exhibit 15.3 shows this variables-by-species table, computed as follows. Take species *a* and variable *Depth* as an example. The relative frequencies of species *a* (i.e., the column profile) are 0, 26/404 = 0.0644, 0, 0, 13/404 = 0.0322, 31/404 = 0.0767, and so on (see Exhibit 1.1). These are used to compute a weighted average of the depth values at each site:

$$0 \times 72 + 0.0644 \times 75 + 0 \times 59 + 0 \times 64 + 0.0322 \times 61 + 0.0767 \times 94 + \cdots = 78.77$$

	a	*b*	*c*	*d*	*e*
Depth	78.77	81.16	72.58	79.27	70.17
Pollution	3.11	3.24	5.49	3.78	3.64
Temperature	3.03	3.06	3.04	3.06	3.11
C	0.26	0.18	0.29	0.25	0.30
S	0.26	0.27	0.46	0.32	0.36
G	0.49	0.56	0.25	0.43	0.34

Exhibit 15.3: *Weighted averages of the environmental variables, using the relative abundances of each species across the samples as weights*

So if the species tends to occur with higher abundance in deeper samples, then the weighted average will be high. Species c, for example, must be occurring in higher abundances in samples with high pollution, and species e in samples of lower depths, which can be verified in Exhibit 15.2. In the case of the dummy variables for sediment, the three values for each species sum to 1 and code the proportion of total abundance of that species in each category.

To obtain the equivalent result as a CCA by analysing the matrix in Exhibit 15.3 needs some technical explanation, since it involves the covariance matrix of the explanatory variables, but the point is that, once the appropriate transformations are made, the inertia in this table is identical to the restricted inertia of 0.249. Knowing this equivalence gives an extra way of thinking about the connection between the species abundances and environmental variables in the triplot.

Coding of explanatory variables

CCA (or the equivalent constrained methods in PCA and LRA) adds the condition that the ordination axes should be linearly related to the explanatory variables. Linearity of the relationship might not be realistic in most circumstances, so just like in regression analysis we can contemplate introducing transformations of the variables, for example, logarithmic transformation, or including polynomial terms, or fuzzy coding. In Chapter 11 we discussed an indirect gradient analysis of the "Barents fish" data set, coding the environmental variables either linearly or fuzzily, and also the geographical position of the samples either in a crisp way in terms of their regional location, or in a fuzzy way based on fuzzy latitude and longitude coordinates. In Chapter 13 various CAs of this same data set were considered. So now we can see how CCA performs on these data, and we will contrast the different ways of coding the environmental variables. Exhibit 15.4 shows the CCA triplot based on linear constraints on the two environmental variables and the 10 dummy variables for the spatial position. Of the total inertia of 2.781 in the abundance data, the environmental variables account for 1.618, that is 58.2%. Of this latter amount, 61.9% is displayed in Exhibit 15.4.

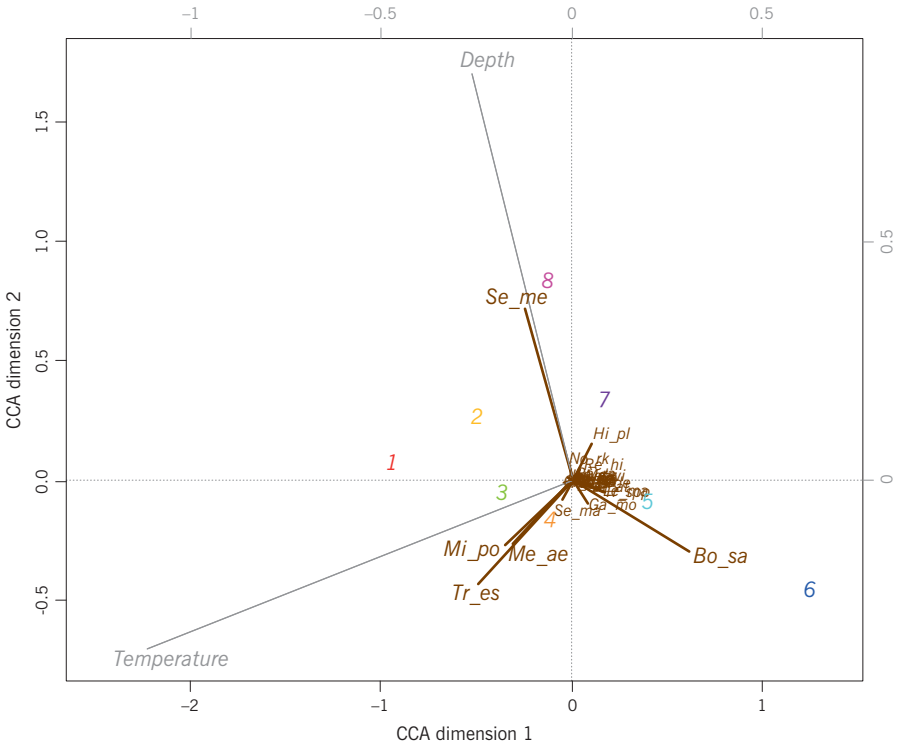


Exhibit 15.4: *CCA of "Barents fish" data, showing highly contributing species in larger font and the two continuous environmental variables according to their regressions on the axes. The 89 sampling sites are not shown, but their averages in the eight regional groupings are*

As a contrast, we try the CCA using depth and temperature coded into four fuzzy categories each, and the spatial position into nine fuzzy categories as explained in Chapter 11, leading to Exhibit 15.5. This time slightly more of the biological-environmental relationship is captured in the restricted space (1.700, i.e. 61.1%, compared to 58.2% previously), which is understandable since nonlinearities in depth and temperature can be picked up thanks to the fuzzy coding. In fact, temperature does appear to have a nonlinear pattern in the triplot, with low and high temperatures being associated with lower values of depth. 60.4% of this constrained ordination is explained in the triplot.

A distinct advantage of Exhibit 15.5 over Exhibit 15.4 is that all the environmental variables, including spatial ones, are displayed in the same way, in this case as weighted averages of the sample positions. To explain this weighted averaging more specifically, Exhibit 15.6 shows the positions of the 89 sample sites corresponding to the CCA result of Exhibit 15.5. Every station has associated with it the values of each fuzzy category, that is 17 values between 0 and 1 inclusive for the 4 fuzzy values of depth, 4 fuzzy values for temperature and 9 fuzzy values for spatial position. For example, for the fuzzy category *d4* the 89 samples have

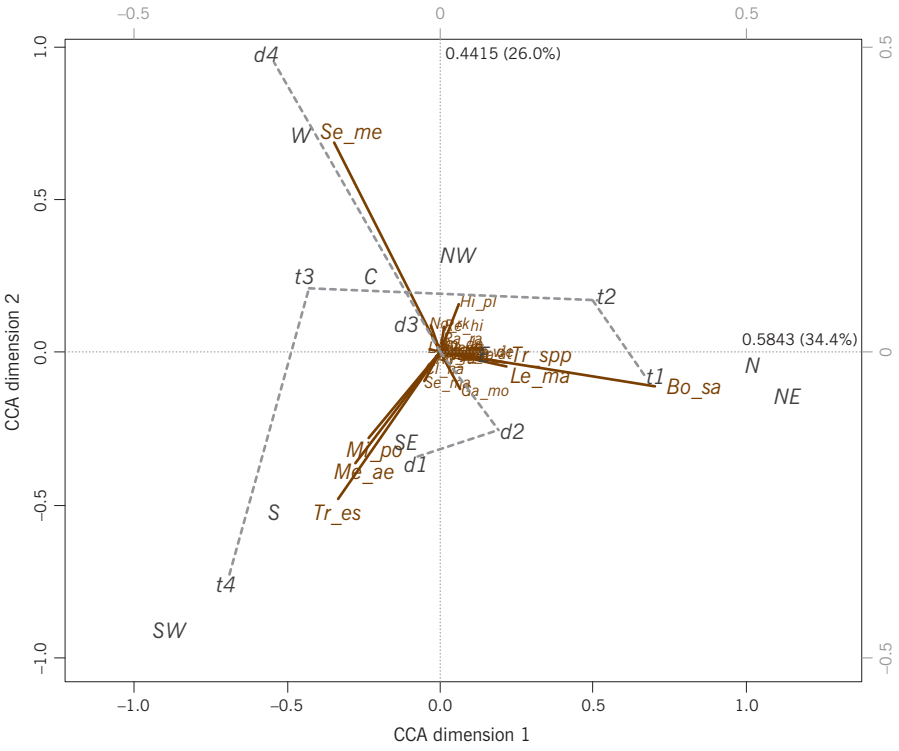


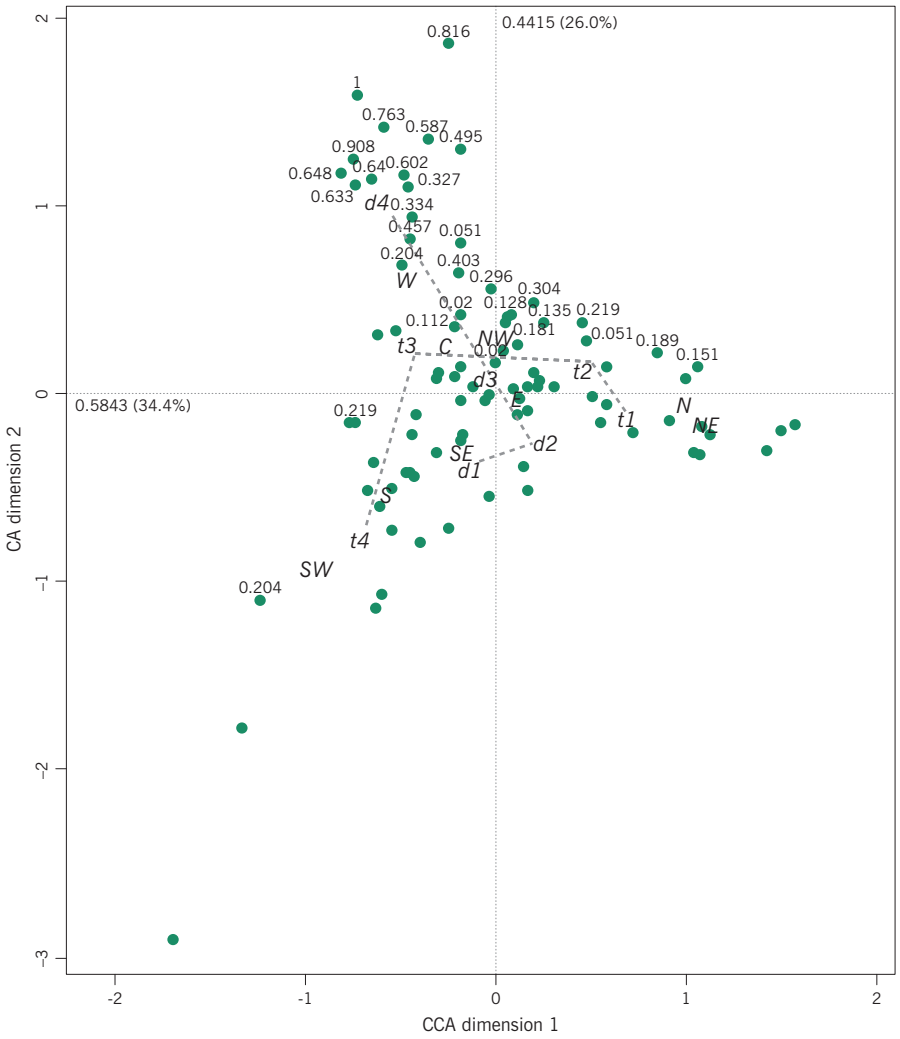
Exhibit 15.5:
CCA triplot of "Barents fish" data, with environmental variables coded into fuzzy categories. Again, sample sites are not shown (see Exhibit 15.6) but the weighted averages of all the fuzzy coded categories are, including the nine fuzzy spatial categories (eight compass points and central category)

59 values equal to 0 and 30 positive values varying from 0.020 to 1.000. These 30 values are shown in Exhibit 15.6 marking their corresponding sample positions, the remaining samples all having zero weights. These 30 samples are almost all in the upper and especially upper left of the ordination and the position of *d4* is at the weighted average of the positions of these 30 samples using those fuzzy values as weights. Especially in the upper left, those marked stations have high values of depth and thus high fuzzy values on category *d4* (the maximum depth is indicated by the value 1), and this leads to *d4* being where it is. In a similar way, all the other fuzzy categories have positions according to the weights placed on the sample points by the respective positive values in the fuzzy coding of the categories.

CCA as a discriminant analysis

The difference between CA of a table of abundances, say, and CCA of the same table constrained by some environmental variables, is that CCA tries to separate the samples on dimensions coinciding with the environmental variation. Thus, in Exhibits 15.5 and 15.6, which one can imagine overlaid to give the triplot of samples, species and variables, separation of the samples is achieved so that the categories of depth, temperature and spatial position are optimally separated in

Exhibit 15.6:
Positions of 89 samples in the CCA of Exhibit 15.5. Each category is at the weighted average of the sample positions, using the fuzzy values as weights. The positive values for category d4 are shown numerically at the respective sample positions



the ordination. If there is a single environmental categorical variable, coded as a set of (crisp) dummy variables, then CCA with that variable as the constraining variable simplifies to a CA that can be thought of as a discriminant analysis between the categories. For example, suppose in the same data set we had a sediment type associated with each sample. Then all the abundances could be aggregated into each sediment type to obtain a sediment-by-species table where the (i,j)-th element would be the total abundance of species j in the sample with sediment type i. The CA of this aggregated table, with all the individual samples as supplementary points, is identical to the CCA with sediment as a categorical constraining variable.

SUMMARY: Canonical correspondence analysis

1. PCA, LRA and CA all have their constrained versions when additional information is available on each biological sample, and these additional variables are considered as predictors of the biological variation.

2. Canonical correspondence analysis (CCA) is the CA of an appropriate table regarded as responses (for example, an abundance matrix) where the dimensions of the result are constrained to be linear combinations of the predictor variables (for example, environmental variables). These predictor variables can be continuous or discrete.

3. CCA projects the response data onto the space of the predictors, and performs a CA in this restricted space. There is thus a splitting of the response inertia into two parts: the part related linearly to the predictors and the part unrelated to the predictors. The inertia of the former part becomes the new total inertia that is decomposed along ordination axes of the CCA. The biological variation that is unrelated to chosen predictors can also be of interest, especially when the variation due to a predictor variable needs to be partialled out of the analysis – this is then called *partial CCA*.

4. There are several advantages of coding the predictor variables fuzzily: nonlinear relationships between the ordination axes and the predictors can be handled, more of the response variable variance is usually explained, and the interpretation of the triplot is unified since all predictors are coded in a categorical way.

5. When there is just one predictor that is discrete, then the CCA constrained by this predictor is equivalent to a CA of the table of response data aggregated into the predictor categories, which in turn is a type of discriminant analysis between these categories.

List of Exhibits

Exhibit 15.1: CA biplot of the biological data in the "bioenv" data set, with samples in principal coordinates and species in contribution coordinates. The one discrete and three continuous environmental variables are shown according to their regression coefficients and the discrete variable's categories are additionally shown (in black) at the centroids of the samples in the corresponding categories 191

Exhibit 15.2: Canonical correspondence analysis triplot of "bioenv" data. The row-principal scaling with species in contribution coordinates is again shown, as well as the environmental variables regressed onto the ordination axes. Percentages of inertia explained are with respect to the restricted inertia 193

Exhibit 15.3: Weighted averages of the environmental variables, using the relative abundances of each species across the samples as weights 195

Exhibit 15.4: CCA of "Barents fish" data, showing highly contributing species in larger font and the two continuous environmental variables according to their regressions on the axes. The 89 sampling sites are not shown, but their averages in the eight regional groupings are 196

Exhibit 15.5: CCA triplot of "Barents fish" data, with environmental variables coded into fuzzy categories. Again, sample sites are not shown (see Exhibit 15.6) but the weighted averages of all the fuzzy coded categories are, including the nine fuzzy spatial categories (eight compass points and central category) 197

Exhibit 15.6: Positions of 89 samples in the CCA of Exhibit 15.5. Each category is at the weighted average of the sample positions, using the fuzzy values as weights. The positive values for category *d4* are shown numerically at the respective sample positions 198