1 Introduction and data types

1.1 Why ordination?

When you investigate the variation of plant or animal communities across a range of different environmental conditions, you typically find not only large differences in species composition of the studied communities, but also a certain consistency or predictability of this variation. For example, if you look at the variation of grassland vegetation in a landscape and describe the plant community composition using vegetation plots, then the individual plots can be usually ordered along one, two or three imaginary axes. The change in the vegetation composition is often small as you move your focus from one plot to those nearby on such a hypothetical axis.

This gradual change in the community composition can often be related to differing, but partially overlapping demands of individual species for environmental factors such as the average soil moisture, its fluctuations throughout the season, the ability of species to compete with other ones for the available nutrients and light, etc. If the axes along which you originally ordered the plots can be identified with a particular environmental factor (such as moisture or richness of soil nutrients), you can call them a soil moisture gradient, or a nutrient availability gradient. Occasionally, such gradients can be identified in a real landscape, e.g. as a spatial gradient along a slope from a riverbank, with gradually decreasing soil moisture. But more often you can identify such axes along which the plant or animal communities vary in a more or less smooth, predictable way, yet you cannot find them in nature as a visible spatial gradient and neither can you identify them uniquely with a particular measurable environmental factor. In such cases, we speak about **gradients of species composition change**.

The variation in biotic communities can be summarised using one of a wide range of statistical methods, but if we stress the continuity of change in community composition, the so-called **ordination methods** are the tools of the trade. They have been used by ecologists since the early 1950s, and during their evolution these methods have radiated into a rich and sometimes confusing mixture of various techniques. Their simplest use can be illustrated by the example introduced above. When you collect data (cases) representing the species composition of selected quadrats in a vegetation stand, you can arrange the cases into a table where individual species are represented by columns and individual cases by rows. When you analyse such data with an ordination method (using the approaches described in this book), you can obtain a fairly representative summary

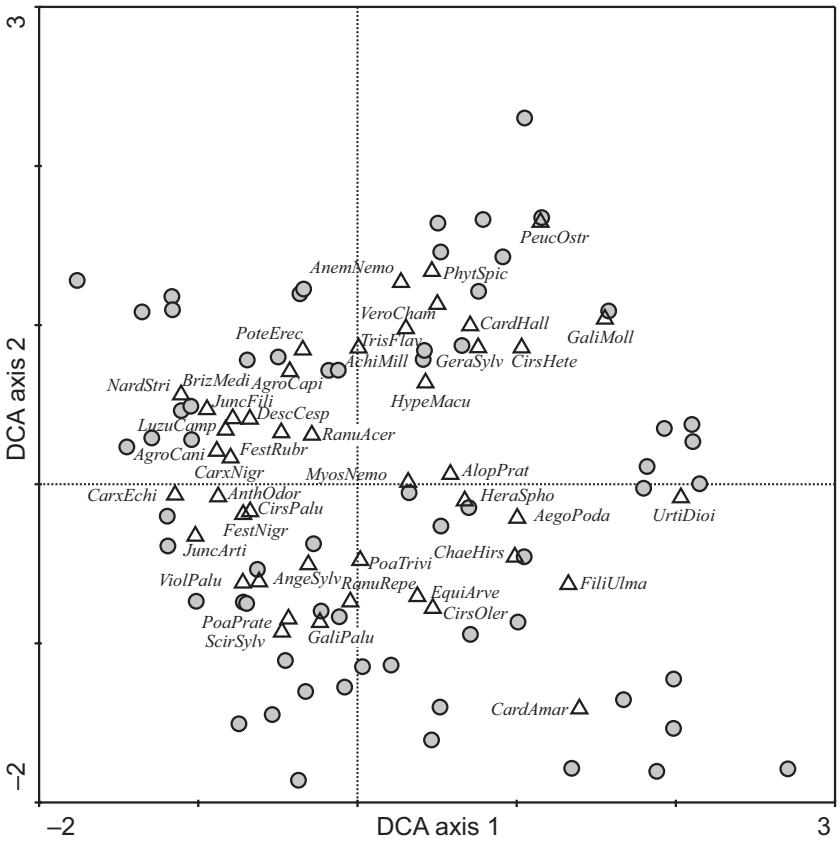

Figure 1–1 Summarising grassland vegetation composition with ordination: ordination diagram from detrended correspondence analysis displaying first two axes, explaining respectively 13% and 8% of the total variation in community composition.

of the grassland vegetation using an ordination diagram, such as the one displayed in Figure 1–1.[1]

The rules for reading such ordination diagrams will be discussed thoroughly later on (in Chapter 11), but even without their knowledge you can read much from the diagram, using the idea of continuous change of composition along the gradients (suggested here by the diagram axes) and the idea that **proximity implies similarity**. The individual cases are represented in Figure 1–1 by grey circles. You can expect that two cases that lay near to each other will be much more similar in their lists of occurring species and

[1] In a research paper, it is appropriate to describe the identity of ordination axes either directly using axes labels (as we do in Figure 1–1) or in figure caption (also illustrated in the caption of Figure 1–1). This book, however, contains a multitude of ordination diagrams and to save some ink, we often omit the labelling of their axes.

even in the relative importance of individual species populations, compared to cases far apart in the diagram.

The triangle symbols represent the individual plant species occurring in the studied type of vegetation (not all species present in the data are shown in the diagram). In this example, our knowledge of the ecological properties of the displayed species[2] can aid us in an **ecological interpretation of the gradients** represented by the diagram axes. The species preferring nutrient rich soils (such as *Urtica dioica*, *Aegopodium podagraria*, or *Filipendula ulmaria*) are located at the right side of the diagram, while the species occurring mostly on soils poor in available nutrients are on the left side (*Viola palustris*, *Carex echinata*, or *Nardus stricta*). The horizontal axis can therefore be informally interpreted as a gradient of nutrient availability, increasing from the left to the right side. Similarly, the species with their points at the bottom of the diagram are from the wetter stands (*Galium palustre*, *Scirpus sylvaticus*, or *Ranunculus repens*) than the species in the upper part of the diagram (such as *Achillea millefolium*, *Trisetum flavescens*, or *Veronica chamaedrys*). The second axis, therefore, represents a gradient of soil moisture.

As you probably already guessed, the proximity of species symbols (triangles) with respect to a particular case symbol (a circle) indicates that these species are likely to occur there more often and/or with a higher (relative) abundance than the species with symbols more distant from the case.

Our example illustrates a frequent way of using ordination methods in community ecology. You can use such an analysis to visually summarise community patterns in an intuitive way and compare the suggested gradients with your independent knowledge of environmental conditions. But you can also test statistically the predictive power of such knowledge; i.e. address the questions such as 'Does the community composition change with the soil moisture or are the identified patterns just a matter of chance?' Such analyses can be done with the help of **constrained ordination methods** and their use will be illustrated later in this book.

However, you do not need to stop with such exploratory or simple confirmatory analyses and this is the focus of the rest of the book. The rich toolbox of various types of regression and analysis of variance, including analysis of repeated measurements on permanent sites, analysis of spatially structured data, various types of hierarchical analysis of variance (ANOVA), etc., allows ecologists to address more complex, and often more realistic questions. Given the fact that the populations of different species occupying the same environment often share similar strategies in relation to the environmental factors, it would be very profitable if one could ask similar complex questions for the whole biotic community. In this book, we demonstrate that this can be done and we show you how to do it.

Unlike the statistical models with a single response variable, the constrained ordination methods enable you to simply compare all the response variables (species) present in

[2] The knowledge of habitat preferences of many plant species is a traditional asset of plant ecologists in Europe. It might be, however, lacking for other groups of organisms or other parts of the world.

the data in terms of their relation to predictors (e.g. environmental variables or human impacts), but also to interpret their similarity and differences using known properties, often representing so-called functional traits of biotic species. And this allows you to relate the functional traits (directly or indirectly) to the properties of environment, generalising your findings beyond the context of the particular area and particular group of organisms you have studied. In this book, we demonstrate the methods working with species traits in sufficient depth for their practical use with Canoco 5.

And yet another type of question arises when you start to compare different kinds of biotic communities. Imagine, for example, that you were able to extend your data set with records on grassland plant community composition with another one, where the community of leaf-eating insect herbivores was quantified for each recorded vegetation plot. How does the compositional variation within the plant and insect communities relate and can we find some gradients, summarising in some optimal way their relation (co-variation)? We will demonstrate a useful method addressing such questions.

1.2 Data types

The terminology for multivariate statistical methods is quite complicated. There are at least two different sets of terms. One, more general and abstract, contains purely statistical terms applicable across the whole field of science. In this chapter we give the terms from this set in italics. The other set contains terms coming from the application domain. As an example, if you study marine phytoplankton, you think about the data in terms of phytoplankton species, sampling station, and environmental characteristics. Starting with version 5, Canoco expects you to define these domain-specific terms (called **item terms** in Canoco 5) for each data table and it uses them afterwards throughout its user interface whenever possible. As this second set varies among projects, we will use the statistical terms in this book, except in the case study chapters or where the discussed concepts are strictly bound to the notion of biological species.

In all multivariate statistical methods we have one data table that can be labelled as the **response data**. This data table contains a collection of observations – *cases*. Each *case* comprises values for multiple *response variables*. The response data can be represented by a rectangular matrix (table), where the rows represent individual *cases* and the columns represent individual *response variables* (species, chemical or physical properties of the water or soil, etc.).[3]

If the response data represent the species composition of a community, we describe the composition using various abundance measures, including counts, frequency estimates, or biomass estimates. Alternatively, we might have information only on the presence or

[3] Note that this arrangement is transposed in comparison with the tables used, for example, in traditional vegetation analyses (phytosociological studies). The classical vegetation tables have individual taxa represented by rows and the columns represent individual records or community types.

absence of species in individual *cases*; such data essentially correspond to the list of species present in each of the cases.

> An important feature of the data types introduced in the preceding paragraph is that summing up the values of individual *response variables* (species) within each *case* results in a meaningful characteristic: for species abundances, this is the total abundance in a case, for presence–absence data (recorded using 1 and 0 values), this is the total number of species in each case. Data tables with this type of values are called **compositional** in Canoco 5 – as opposed to the **general** type – and this is an important attribute you must set correctly for each created data table to obtain useful advice for your analyses.

In some cases, we estimate the values for the response data on a simple, semi-quantitative scale. Good examples are the various semi-quantitative scales used in recording the composition of plant communities (e.g. original Braun-Blanquet scale or its various modifications).[4]

If our response variables represent the properties of the chemical or physical environment (e.g. concentrations of ions or more complicated compounds in the water, soil acidity, water temperature, etc.), we usually get quantitative values for them, but with an additional limitation: these characteristics do not share the same units of measurement and cannot be meaningfully added, even if they share the units (such as $\mu g \cdot l^{-1}$ for various ion types[5]). In other words, these are non-compositional, *general* data as discussed above. This fact precludes the use of some of the ordination methods[6] and dictates the way the variables are standardised if used in the other ordinations (see Section 1.3).

Very often the response data table is accompanied by another one containing predictor variables that we want to use to understand the response data table contents. If our response data represent community composition, then the predictor data set typically contains measurements of the soil or water properties (for the terrestrial or aquatic ecosystems, respectively), a semi-quantitative or categorical scoring of human impact, etc. When using these variables in a model to predict the response data (like community composition), we might divide them into two different types. The first type is called the *explanatory variables* and refers to the variables that are of the prime interest (in the role of predictors) in our particular analysis. The other type represents the *covariates* which are also variables with an acknowledged (or hypothesised) influence on the

[4] And although the sums of such values are sometimes not fully intuitive entities, they still represent rough estimates of the more precise abundance estimates and so we should treat them as compositional data type too.

[5] We admit that in some context, adding concentrations of various ions makes sense, but in ecological studies, these (additive) concentrations are usually supplemented with other chemical or physical measures that are not additive.

[6] Namely correspondence analysis (CA), detrended correspondence analysis (DCA), or canonical correspondence analysis (CCA) and related partial versions. But such general variables can be used in these methods as supplementary or explanatory variables or as covariates – they only cannot be used as the response data for them.

response variables. We want to account for (factor-out or partial-out) such an influence **before** focusing on the influence of the variables of prime interest (i.e. on the effect of *explanatory* variables).

As an example, imagine a situation where you study the effects of soil properties and type of management (hay cutting or pasturing) on the species composition of grassland in a particular area. In one analysis, you might be interested in the effect of soil properties, paying no attention to the management regime. In this analysis, you use the grassland composition as the *response data* (with individual plant species as individual *response variables*) and the measured soil properties as the *explanatory variables*. Based on the results, you can make conclusions about the relation of plant species populations to particular environmental gradients, which are described (more or less appropriately) by the measured soil properties. Similarly, you can ask how the management type influences plant composition. In the corresponding analysis, the variables describing the management regime act as *explanatory variables*. Further, you might expect that the management also influences the soil properties and this is probably one of the ways in which management acts upon the community composition. Based on such expectation, you may ask about the influence of management regime **beyond** that mediated through the changes of soil properties. To address such a question, you must use the variables describing the management regime as the *explanatory variables* and the measured soil properties as the *covariates*.[7] Of course, there might also exist unique effects of soil properties not related to management, and to test and explore them you need to define another analysis, where the management descriptors act as covariates and the soil characteristics as explanatory variables.

Another typical example of *covariate* use is for an experimental design where *cases* are grouped into logical or physical blocks. The values of *response variables* (e.g. species composition) for a group of *cases* might be similar due to their (spatial) proximity, so we need to model this influence and account for it in our data. The differences in *response variables* that are due to the membership of *cases* in different blocks can be removed (i.e. 'partialled-out') from the model by using a factor identifying experimental blocks as a *covariate*.

Beside *explanatory variables* and *covariates*, we recognise yet another kind of predictor variables, called *supplementary variables*. These are used in unconstrained ordination (also called indirect gradient analysis), defined in the following section, to interpret its results.[8]

Predictors can be quantitative variables (concentration of nitrate ions in soil), semi-quantitative estimates (degree of human influence estimated on a 0–3 scale) or factors (nominal or categorical – also categorial – variables). The simplest predictor form is a binary variable, where the presence or absence of a certain feature or event (e.g. vegetation was mown, trap is located near a road, etc.) is indicated, respectively, by a 1 or 0 value.

[7] This particular example is discussed in Canoco 5 manual, section 6.3.1.
[8] Supplementary variables are projected *post hoc* into an ordination space already computed for cases and response variables.

Factors with multiple values (levels) are the natural way of expressing the classification of our *cases* or subjects: for example, classes of management type for meadows or the type of stream for a study of pollution impact on rivers.

1.3 Data transformation and standardisation

1.3.1 Transformation

As will be shown in Chapter 4, ordination methods find the axes representing regression predictors that are optimal for predicting the values of *response variables*, i.e. the values in the response data table. Therefore, the problem of selecting a transformation for the *response variables* is rather similar to the problem one needs to solve when using any of the variables in the (multiple) regression. The one additional restriction is the need to specify an identical data transformation for all the *response variables* when working with so-called compositional data, see the preceding section, because such variables are often measured on the same scale. In the unimodal (weighted averaging) ordination methods (see Section 4.2), the data values cannot be negative and this imposes a further restriction on the outcome of any potential transformation.

This restriction is particularly important in the case of the **log transformation**. The logarithm of 1.0 is zero and logarithms of values between 0 and 1 are negative and hence non-acceptable for unimodal ordination. Therefore, Canoco provides a flexible log-transformation formula

$$y' = \log(A \cdot y + B)$$

You should specify the values of A and B so that before the transformation is applied to your data values, the result $A \cdot y + B$ is always greater than zero. The default values of both A and B are 1.0, which neatly map the zero values again to zero, and positive y values remain positive. Nevertheless, if your original values are small (say, in the range 0.0 to 0.1), the shift caused by adding the relatively large value of 1.0 dominates the resulting structure of the data matrix. You can adjust the transformation in this case by increasing the value of A to 10.[9] The default log transformation (i.e. $\log(y + 1)$) works well with the percentage data on the 0 to 100 scale, or with the ordinary counts of objects (e.g. caught individuals of each species).

Whether to use a log transformation or keep the original scale is a difficult question, with different answers from different statisticians. We suggest that you do not consider the variable distribution (at least not in the sense of testing its difference from a Normal distribution, as routinely and often incorrectly done[10]), but you base your decision on how you phrase the hypothesis standing behind your research, as described in the following paragraph.

[9] Such change is automatically done by Canoco when you, for example, specify A = 1 and B = 0.1, changing A to 10.0 and B to 1.0.

[10] Many users forget that only the residuals of a statistical model are expected to have a Normal distribution and they test the response variable values instead.

As stated above, ordination methods can be viewed as an extension of multiple regression methods, so this **semantic-based approach** will be explained in the simpler regression context. You might try to predict the abundance of a particular biotic species in cases, based on the values of one or more predictors (environmental variables, or ordination axes in the context of ordination methods). One can formulate the question addressed by such a regression model (assuming just a single predictor variable for simplicity) as 'How does the average value of species Y change with a change in the environmental variable X by one unit?' If neither the response variable nor the predictors are log transformed, the answer coming from a regression model can take the form: 'The value of species Y increases by B if the value of environmental variable X increases by one measurement unit'. Of course, B is then the regression coefficient of the linear model equation $Y = B_0 + B \cdot X + E$. But often you can feel that the answer should have a different form, such as 'If the value of environmental variable X increases by one unit, the average abundance of the species increases by 10%.' Alternatively, you can say, 'the abundance increases 1.10 times'. In both cases, you are thinking on a multiplicative scale, which is not the scale assumed by the linear regression model. In such a situation, you should **log-transform the response variable**. Similarly, if the effect of a predictor (environmental) variable changes in a multiplicative way, **the predictor variable should be log-transformed**.[11]

Plant community composition data are often collected on a semi-quantitative estimation scale and the Braun-Blanquet scale with seven levels (r, $+$, 1, 2, 3, 4, 5) is a typical example. Such a scale is then quantified in the spreadsheets using corresponding ordinal levels (from 1 to 7 in the above case). This coding, called **ordinal transformation**, already implies a log-like transformation because the actual cover/abundance differences between the successive levels are generally increasing. An alternative approach to using such estimates in data analysis is to replace them by the assumed centres of the corresponding range of percentage cover. But doing so, however, you find a problem with the r and $+$ levels because these are based more on the abundance (number of individuals) of the species than on their estimated cover. Nevertheless, using very rough replacements, such as 0.1 for r and 0.5 for $+$, rarely harms the analysis (compared to the alternative solutions).

Another useful transformation of the response data available in Canoco is the square-root transformation. This might be the best transformation to apply to count data, such as the number of specimens of individual species collected in a soil trap, number of individuals of various ant species passing over a marked 'count line', etc., but the log transformation also handles well such data. Further transformations available in Canoco 5 analyses are two variants of arcsine transformation (one for fractional data on a 0–1 scale, another one for percentage data on the 0–100 scale) and binarising transformation, turning any positive value into 1.0 and other values into 0.0. Additionally,

[11] Strictly speaking, the log-transformation turns a multiplicative relation into an additive one only if you can use it without the B constant (see the formula above). But even if the presence of zero values in your data requires you to add a positive B, the bias is small if B is small compared with the range of transformed variable values.

if you need any kind of transformation that is not provided by the Canoco software, you might do it in your spreadsheet software and import the transformed data into Canoco project.

When you work with binary (0/1) data, any of the transformations discussed above do not have any real effect, so it is best to keep such data untransformed.

The transformation you choose for a data table when it is used as the response data for the first time is remembered and offered during the setup of following analyses. You can also set or change the default transformation directly using the *Data | Default transformation and standardization* menu command, when the particular table is in foreground. Shared transformation can be set only for data tables of the *compositional* type (see Section 1.2), but for the *general* table type you can set the implicit transformation for individual variables in the table.

1.3.2 Standardisation

In this book, we treat the transformation and standardisation processes separately, even though both 'transform' (change) the original data in the usual meaning. In our view, the **transformation** can be represented by an algebraic function $Y'_{ik} = f(Y_{ik})$ which is applied to each value independently of the other values. **Standardisation** is done, on the other hand, with respect to either the values of other variables measured for the same case (standardisation by cases) or the values of the same variable measured for the other cases (standardisation by variables).

In fact, even the term standardisation can be understood more broadly, as we do here, or in a more narrow sense, as used in the Canoco software: *standardisation* means there the adjustment of values affecting their variability, while so-called **centring**[12] changes mean value. The most common type of centring leads to zero average of variables (or – more rarely used – of cases), while the most common type of standardisation (in the narrower sense) is the standardisation to unit norm (a square root of the sum of squared variable/case values).[13] For variables, the standardisation to unit norm must be almost always combined with the centring, so that the resulting variables have not only a unit norm, but also a unit variance.

Canoco centres and standardises any *explanatory* or *supplementary variables* and any *covariates*, to bring their means to zero and their variances to one,[14] but for the *response variables*, use of standardisation (and of centring, to a lesser extent) is an important choice in the **linear** ordination methods.[15] For constrained linear methods (i.e. redundancy analysis), the centring by variables is required (and enforced in Canoco 5),

[12] Note that the Canoco 5 user interface uses US spelling ('center'/'centering') and so we do the same whenever referring directly to program user interface elements.

[13] Alternatively, the Analysis Setup Wizard offers the standardisation of cases to unit sum, but see Table 6–1 and Section 6.2.2 for a discussion of existing issues with this standardisation.

[14] This treatment of predictors makes their effects, as seen in ordination diagrams and in numerical summaries, comparable, but it also assures numerical stability of the calculations.

[15] Not so in unimodal methods, where a special form of double standardisation (both by rows and by variables) is implied by the weighted averaging algorithm and the standardisation cannot be therefore selectively applied by the user.

while the standardisation is optional (also in unconstrained linear ordination), at least for compositional data tables.

> With compositional data tables, you should be extremely careful with standardisation by variables (typically species). The intention of this procedure is to give all the species (response variables) the same weight. But the result is often counterproductive, because a species with a low frequency of occurrence might become very influential. If the species is found in one case only, then all of its quantity is in this case alone, which makes this case very different from the others. On the other hand, the species that are found in many cases do not attain, after standardisation, a high share in any of them and their effect is relatively small.

For general (non-compositional) data tables where each variable has its own scale, it is necessary to centre and standardise the variables (this is also often referred to as calculating the *z-scores*). A typical example of this comes from classical taxonomy: each object (individual, population) is described by several characteristics, measured in different units (e.g. number of petals, density of hairs, weight of seeds, etc.). When a similarity among measured individuals or among populations is calculated from the rough data, the weight of individual variables changes when you change their units – and the final result is completely meaningless.

The difference of running principal components analysis (PCA) on response data that were only centred or both centred and standardised by variables is reflected in the traditional names for these two variants: the former one is called 'PCA on variance-covariance matrix', while the latter one is called 'PCA on correlation matrix'.[16]

For the redundancy analysis (RDA, constrained linear ordination), Canoco also offers another kind of standardisation by response variables, called the **standardisation by error variance**. In this case, Canoco proceeds as if the standard centring and standardisation by variables was chosen, but in addition it calculates, separately for each response variable, how much of its variance was not explained by the explanatory variables (and covariates, for partial RDA). The inverse of that *error variance* is then used as the relative weight of each response variable. Therefore, the better a response variable is described by the explanatory variables, the greater impact it has on the analysis results.

> For response data representing biotic communities, the standardisation by cases (either by case norm or by the total) has a clear ecological meaning. If you use it, you are interested only in proportions of species (both for the standardisation by totals and by the case norm). With standardisation, two cases containing three species, in the first case with 50, 20 and 10 individuals, and in the second case with 5, 2 and 1 individual, will be found identical.[17] Standardisation by the total (i.e. to percentages)

[16] PCA on correlation matrix is the most common type of PCA outside the field of ecology.

[17] The differences in ecological interpretations of analyses with and without standardisation by case norm are discussed in Section 15.3.