Multivariate Analysis of Ecological Data using CANOCO

JAN LEPŠ
University of South Bohemia, and
Czech Academy of Sciences,
Czech Republic

PETR ŠMILAUER
University of South Bohemia,
Czech Republic

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typefaces Lexicon No. 2 10/14 pt and Lexicon No. 1 *System* LATEX 2ε [TB]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Lepš, Jan 1953–
Multivariate analysis of ecolological data using CANOCO / Jan Lepš & Petr Šmilauer.
p. cm.
Includes bibliographical references (p.).
ISBN 0 521 81409 X (hb.) ISBN 0 521 89108 6 (pb.)
1. Ecology – Statistical methods. 2. Multivariate analysis. I. Šmilauer, Petr, 1967-II.
Title.

QH541.15.S72L47 2003
577′.07′27–dc21 2002034801

ISBN 0 521 81409 X hardback
ISBN 0 521 89108 6 paperback

Contents

1

Introduction and data manipulation

1.1. Why ordination?

When we investigate variation of plant or animal communities across a range of different environmental conditions, we usually 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 we look at the variation of grassland vegetation in a landscape and describe the plant community composition using vegetation samples, then the individual samples can be usually ordered along one, two or three imaginary axes. The change in the vegetation composition is often small as we move our focus from one sample 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 we originally ordered the samples can be identified with a particular environmental factor (such as moisture or richness of soil nutrients), we can call them a soil moisture gradient, a nutrient availability gradient, etc. 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 we can identify such axes along which the plant or animal communities vary in a more or less smooth, predictable way, yet we cannot find them in nature as a visible spatial gradient and neither can we 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 summarized using one of a wide range of statistical methods, but if we stress the continuity of change

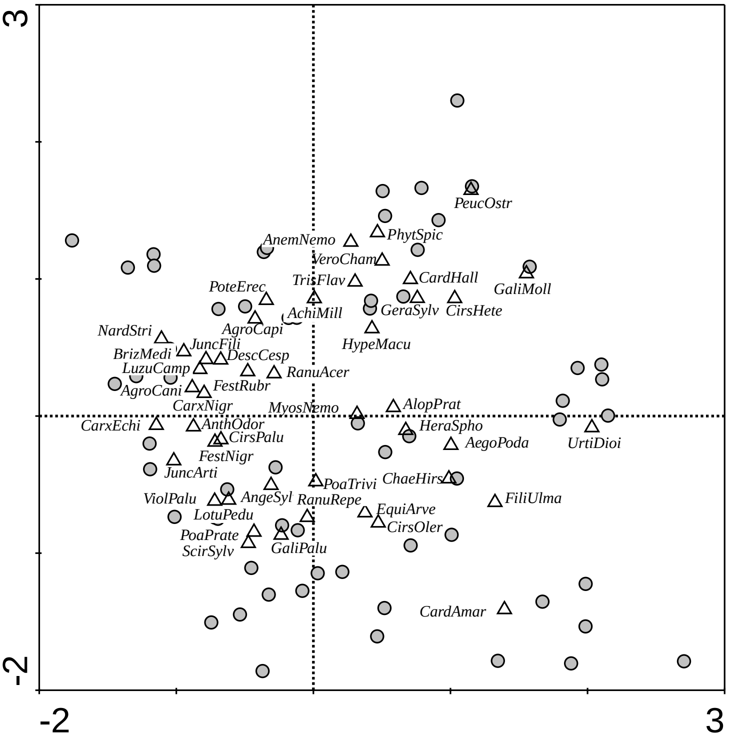


Figure 1-1. Summarizing grassland vegetation composition with ordination: ordination diagram from correspondence analysis.

in community composition, the so-called **ordination methods** are the tools of 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 we collect recordings (samples) representing the species composition of a selected quadrat in a vegetation stand, we can arrange the samples into a table where individual species are represented by columns and individual samples by rows. When we analyse such data with an ordination method (using the approaches described in this book), we can obtain a fairly representative summary of the grassland vegetation using an ordination diagram, such as the one displayed in Figure 1-1.

The rules for reading such ordination diagrams will be discussed thoroughly later on (see Chapter 10), but even without their knowledge we 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 samples are represented in Figure 1-1 by grey circles. We can expect that two samples that lie near to each other will be much more similar in terms of list of occurring species and even in the relative importance of individual species populations, compared to samples 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 were included in the diagram). In this example, our knowledge of the ecological properties of the displayed species 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 in 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 have probably already guessed, the proximity of species symbols (triangles) with respect to a particular sample symbol (a circle) indicates that these species are likely to occur more often and/or with a higher (relative) abundance than the species with symbols more distant from the sample.

Our example study illustrates the most frequent use of ordination methods in community ecology. We can use such an analysis to summarize community patterns and compare the suggested gradients with our independent knowledge of environmental conditions. But we 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?' These analyses can be done with the help of **constrained ordination methods** and their use will be illustrated later in this book.

However, we 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 communities. In this book, we demonstrate that this can be done and we show the reader how to do it.

1.2. Terminology

The terminology for multivariate statistical methods is quite complicated. There are at least two different sets of terminology. One, more general and abstract, contains purely statistical terms applicable across the whole field of science. In this section we give the terms from this set in italics and mostly in parentheses. The other represents a mixture of terms used in ecological statistics with the most typical examples coming from the field of community ecology. This is the set on which we will focus, using the former just to refer to the more general statistical theory. In this way, we use the same terminology as the CANOCO software documentation.

In all cases, we have a data set with the **primary data**. This data set contains records on a collection of observations – **samples** (*sampling units*).[*] Each sample comprises values for multiple **species** or, less often, the other kinds of descriptors. The primary data can be represented by a rectangular matrix, where the rows typically represent individual samples and the columns represent individual variables (species, chemical or physical properties of the water or soil, etc.).[†]

Very often our primary data set (containing the *response variables*) is accompanied by another data set containing the *explanatory variables*. If our primary data represent community composition, then the explanatory data set typically contains measurements of the soil or water properties (for the terrestrial or aquatic ecosystems, respectively), a semi-quantitative scoring of human impact, etc. When we use the *explanatory variables* in a model to predict the primary data (like community composition), we might divide them into two different groups. The first group is called, somewhat inappropriately, the **environmental variables** and refers to the variables that are of prime interest (in the role of predictors) in our particular analysis. The other group represents the **covariables** (often referred to as *covariates* in other statistical approaches), which are

[*] There is an inconsistency in the terminology: in classical statistical terminology, **sample** means a collection of sampling units, usually selected at random from the population. In community ecology, sample is usually used for a description of a sampling unit. This usage will be followed in this text. The general statistical packages use the term **case** with the same meaning.

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

also explanatory variables with an acknowledged (or hypothesized) influence on the *response variables*. We want to account for (subtract, partial-out) such an influence **before** focusing on the influence of the variables of prime interest (i.e. the effect of environmental variables).

As an example, let us imagine a situation where we study the effects of soil properties and type of management (hay cutting or pasturing) on the species composition of meadows in a particular area. In one analysis, we might be interested in the effect of soil properties, paying no attention to the management regime. In this analysis, we use the grassland composition as the **species data** (i.e. *primary data set*, with individual plant species as individual *response variables*) and the measured soil properties as the **environmental variables** (*explanatory variables*). Based on the results, we can make conclusions about the preferences of individual plant species' populations for particular environmental gradients, which are described (more or less appropriately) by the measured soil properties. Similarly, we can ask how the management type influences plant composition. In this case, the variables describing the management regime act as environmental variables. Naturally, we 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, we may ask about the influence of the management regime **beyond** that mediated through the changes of soil properties. To address such a question, we use the variables describing the management regime as the **environmental variables** and the measured soil properties as the **covariables.**[*]

One of the keys to understanding the terminology used by the CANOCO program is to realize that the data referred to by CANOCO as the **species data** might, in fact, be any kind of data with variables whose values we want to **predict.** For example, if we would like to predict the quantities of various metal ions in river water based on the landscape composition in the catchment area, then the individual ions would represent the individual 'species' in CANOCO terminology. If the **species data** really represent the species composition of a community, we describe the composition using various abundance measures, including counts, frequency estimates, and biomass estimates. Alternatively, we might have information only on the presence or absence of species in individual samples. The quantitative and presence-absence variables may also occur as *explanatory variables*. These various kinds of data values are treated in more detail later in this chapter.

[*] This particular example is discussed in the Canoco for Windows manual (Ter Braak & Šmilauer, 2002), section 8.3.1.

Table 1-1. *The types of the statistical models*

Response	Predictor(s)	
variable(s) . . .	Absent	Present
. . . **is one**	• distribution summary	• regression models *sensu lato*
. . . **are many**	• indirect gradient analysis (PCA, DCA, NMDS) • cluster analysis	• direct gradient analysis • discriminant analysis (CVA)

CVA, canonical variate analysis; DCA, detrended correspondence analysis; NMDS, non-metric multidimensional scaling; PCA, principal components analysis.

1.3. Types of analyses

If we try to describe the behaviour of one or more response variables, the appropriate statistical modelling methodology depends on whether we study each of the response variables separately (or many variables at the same time), and whether we have any explanatory variables (predictors) available when we build the model.

Table 1-1 summarizes the most important statistical methodologies used in these different situations.

If we look at a single response variable and there are no predictors available, then we can only summarize the distributional properties of that variable (e.g. by a histogram, median, standard deviation, inter-quartile range, etc.). In the case of multivariate data, we might use either the ordination approach represented by the methods of **indirect gradient analysis** (most prominent are the principal components analysis – PCA, correspondence analysis – CA, detrended correspondence analysis – DCA, and non-metric multidimensional scaling – NMDS) or we can try to (hierarchically) divide our set of samples into compact distinct groups (methods of cluster analysis, see Chapter 7).

If we have one or more predictors available and we describe values of a single variable, then we use **regression models** in the broad sense, i.e. including both traditional regression methods and methods of analysis of variance (ANOVA) and analysis of covariance (ANOCOV). This group of methods is unified under the so-called **general linear model** and was recently extended and enhanced by the methodology of **generalized linear models (GLM)** and **generalized additive models (GAM)**. Further information on these models is provided in Chapter 8.

If we have predictors for a set of response variables, we can summarize relations between multiple response variables (typically biological species) and one or several predictors using the methods of **direct gradient analysis** (most prominent are redundancy analysis (RDA) and canonical correspondence analysis (CCA), but there are several other methods in this category).

1.4. Response variables

The data table with response variables* is always part of multivariate analyses. If explanatory variables (see Section 1.5), which may explain the values of the response variables, were not measured, the statistical methods can try to construct hypothetical explanatory variables (groups or gradients).

The response variables (often called species data, based on the typical context of biological community data) can often be measured in a precise (quantitative) way. Examples are the dry weight of the above-ground biomass of plant species, counts of specimens of individual insect species falling into soil traps, or the percentage cover of individual vegetation types in a particular landscape. We can compare different values not only by using the 'greater-than', 'less-than' or 'equal to' expressions, but also using their ratios ('this value is two times higher than the other one').

In other cases, we estimate the values for the primary 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). The simplest possible form of data are binary (also called presence-absence or 0/1) data. These data essentially correspond to the list of species present in each of the samples.

If our response variables represent the properties of the chemical or physical environment (e.g. quantified 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 constraint: these characteristics do not share the same units of measurement. This fact precludes the use of some of the ordination methods† and dictates the way the variables are standardized if used in the other ordinations (see Section 4.4).

1.5. Explanatory variables

The explanatory variables (also called predictors or independent variables) represent the knowledge that we have about our samples and that we can use to predict the values of the response variables (e.g. abundance of various

[*] also called dependent variables.

[†] namely correspondence analysis (CA), detrended correspondence analysis (DCA), or canonical correspondence analysis (CCA).

species) in a particular situation. For example, we might try to predict the composition of a plant community based on the soil properties and the type of land management. Note that usually the primary task is not the prediction itself. We try to use ‘prediction rules’ (derived, most often, from the ordination diagrams) to learn more about the studied organisms or systems.

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, the sample is located in study area X, etc.) is indicated, respectively, by a 1 or 0 value.

The factors are the natural way of expressing the classification of our samples or subjects: For example, classes of management type for meadows, type of stream for a study of pollution impact on rivers, or an indicator of the presence/absence of a settlement near the sample in question. When using factors in the CANOCO program, we must re-code them into so-called **dummy variables**, sometimes also called **indicator variables** (and, also, binary variables). There is one separate dummy variable for each different value (level) of the factor. If a sample (observation) has a particular value of the factor, then the corresponding dummy variable has the value 1.0 for this sample, and the other dummy variables have a value of 0.0 for the same sample. For example, we might record for each of our samples of grassland vegetation whether it is a pasture, meadow, or abandoned grassland. We need three dummy variables for recording such a factor and their respective values for a meadow are 0.0, 1.0, and 0.0.*

Additionally, this explicit decomposition of factors into dummy variables allows us to create so-called **fuzzy coding**. Using our previous example, we might include in our data set a site that had been used as a hay-cut meadow until the previous year, but was used as pasture in the current year. We can reasonably expect that both types of management influenced the present composition of the plant community. Therefore, we would give values larger than 0.0 and less than 1.0 for both the first and second dummy variables. The important restriction here is that the values must sum to 1.0 (similar to the dummy variables coding normal factors). Unless we can quantify the relative importance of the two management types acting on this site, our best guess is to use values 0.5, 0.5, and 0.0.

* In fact, we need only two (generally $K-1$) dummy variables to code uniquely a factor with three (generally K) levels. But the one redundant dummy variable is usually kept in the data, which is advantageous when visualizing the results in ordination diagrams.

If we build a model where we try to predict values of the response variables ('species data') using the explanatory variables ('environmental data'), we often encounter a situation where some of the explanatory variables affect the species data, yet these variables are treated differently: we do not want to interpret their effect, but only want to take this effect into account when judging the effects of the other variables. We call these variables **covariables** (or, alternatively, **covariates**). A typical example is an experimental design where samples are grouped into logical or physical blocks. The values of response variables (e.g. species composition) for a group of samples 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 samples in different blocks must be removed (i.e. 'partialled-out') from the model.

But, in fact, almost any explanatory variable can take the role of a covariable. For example, in a project where the effect of management type on butterfly community composition is studied, we might have the localities at different altitudes. The altitude might have an important influence on the butterfly communities, but in this situation we are primarily interested in the management effects. If we remove the effect of the altitude, we might get a clearer picture of the influence that the management regime has on the butterfly populations.

1.6. Handling missing values in data

Whatever precautions we take, we are often not able to collect all the data values we need: a soil sample sent to a regional lab gets lost, we forget to fill in a particular slot in our data collection sheet, etc.

Most often, we cannot go back and fill in the empty slots, usually because the subjects we study change in time. We can attempt to leave those slots empty, but this is often not the best decision. For example, when recording sparse community data (we might have a pool of, say, 300 species, but the average number of species per sample is much lower), we interpret the empty cells in a spreadsheet as absences, i.e. zero values. But the absence of a species is very different from the situation where we simply forgot to look for this species! Some statistical programs provide a notion of missing values (it might be represented as a word 'NA', for example), but this is only a notational convenience. The actual statistical method must deal further with the fact that there are missing values in the data. Here are few options we might consider:

1. We can remove the samples in which the missing values occur. This works well if the missing values are concentrated in a few samples. If we have,

for example, a data set with 30 variables and 500 samples and there are 20 missing values from only three samples, it might be wise to remove these three samples from our data before the analysis. This strategy is often used by general statistical packages and it is usually called 'case-wise deletion'.

2. On the other hand, if the missing values are concentrated in a few variables that are not deemed critical, we might remove the variables from our data set. Such a situation often occurs when we are dealing with data representing chemical analyses. If 'every thinkable' cation concentration was measured, there is usually a strong correlation among them. For example, if we know the values of cadmium concentration in air deposits, we can usually predict the concentration of mercury with reasonable precision (although this depends on the type of pollution source). Strong correlation between these two characteristics implies that we can make good predictions with only one of these variables. So, if we have a lot of missing values in cadmium concentrations, it might be best to drop this variable from our data.
3. The two methods of handling missing values described above might seem rather crude, because we lose so much of our data that we often collected at considerable expense. Indeed, there are various **imputation** methods. The simplest one is to take the average value of the variable (calculated, of course, only from the samples where the value is not missing) and replace the missing values with it. Another, more sophisticated one, is to build a (multiple) regression model, using the samples with no missing values, to predict the missing value of a variable for samples where the values of the other variables (predictors in the regression model) are not missing. This way, we might fill in all the holes in our data table, without deleting any samples or variables. Yet, we are deceiving ourselves – we only duplicate the information we have. The degrees of freedom we lost initially cannot be recovered.

If we then use such supplemented data in a statistical test, this test makes an erroneous assumption about the number of degrees of freedom (number of independent observations in our data) that support the conclusion made. Therefore, the significance level estimates are not quite correct (they are 'over-optimistic'). We can alleviate this problem partially by decreasing the statistical weight for the samples where missing values were estimated using one or another method. The calculation can be quite simple: in a data set with 20 variables, a sample with missing values replaced for five variables gets a weight 0.75 (=1.00 − 5/20). Nevertheless, this solution is not perfect. If we work only with a subset of the variables (for example, during a stepwise selection of explanatory variables), the samples with any variable being imputed carry the penalty even if the imputed variables are not used.

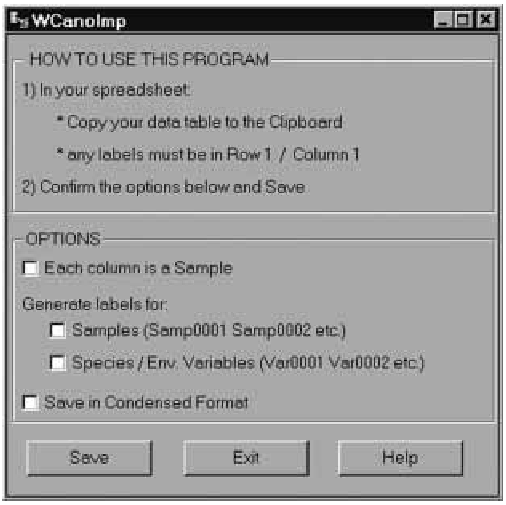
The methods of handling missing data values are treated in detail in a book by Little & Rubin (1987).

1.7. Importing data from spreadsheets – WCanoImp program

The preparation of input data for multivariate analyses has always been the biggest obstacle to their effective use. In the older versions of the CANOCO program, one had to understand the overly complicated and unforgiving format of the data files, which was based on the requirements of the FORTRAN programming language used to create the CANOCO program. Version 4 of CANOCO alleviates this problem by two alternative means. First, there is now a simple format with minimum requirements for the file contents (the free format). Second, and probably more important, is the new, easy method of transforming data stored in spreadsheets into CANOCO format files. In this section, we will demonstrate how to use the WCanoImp program for this purpose.

Let us start with the data in your spreadsheet program. While the majority of users will work with Microsoft Excel, the described procedure is applicable to any other spreadsheet program running under Microsoft Windows. If the data are stored in a relational database (Oracle, FoxBASE, Access, etc.) you can use the facilities of your spreadsheet program to first import the data into it. In the spreadsheet, you must arrange your data into a rectangular structure, as laid out by the spreadsheet grid. In the default layout, the individual samples correspond to the rows while the individual spreadsheet columns represent the variables. In addition, you have a simple heading for both rows and columns: the first row (except the empty upper left corner cell) contains the names of variables, while the first column contains the names of the individual samples. Use of heading(s) is optional, because WCanoImp is able to generate simple names there. When using the heading row and/or column, you must observe the limitations imposed by the CANOCO program. The names cannot have more than eight characters and the character set is somewhat limited: the safest strategy is to use only the basic English letters, digits, dot, hyphen and space. Nevertheless, WCanoImp replaces any prohibited characters by a dot and also shortens any names longer than the eight characters. Uniqueness (and interpretability) of the names can be lost in such a case, so it is better to take this limitation into account when initially creating the names.

The remaining cells of the spreadsheet must only be numbers (whole or decimal) or they must be empty. No coding using other kinds of characters is



HOW TO USE THIS PROGRAM

1) In your spreadsheet:

* Copy your data table to the Clipboard

* any labels must be in Row 1 / Column 1

2) Confirm the options below and Save

OPTIONS

☐ Each column is a Sample

Generate labels for:

☐ Samples (Samp0001 Samp0002 etc.)

☐ Species / Env. Variables (Var0001 Var0002 etc.)

☐ Save in Condensed Format

Figure 1-2. The main window of the WCanoImp program.

allowed. Qualitative variables (‘factors’) must be coded for the CANOCO program using a set of ‘dummy variables’ – see Section 1.5 for more details.

When the data matrix is ready in the spreadsheet program, you must select the rectangular region (e.g. using the mouse pointer) and copy its contents to the Windows Clipboard. WCanoImp takes the data from the Clipboard, determines their properties (range of values, number of decimal digits, etc.) and allows you to create a new data file containing these values, and conforming to one of two possible CANOCO data file formats. Hopefully it is clear that the requirements concerning the format of the data in a spreadsheet program apply only to the rectangle being copied to the Clipboard. Outside of it, you can place whatever values, graphs or objects you like.

The WCanoImp program is accessible from the Canoco for Windows program menu (**Start** > **Programs** > [*Canoco for Windows folder*]). This import utility has an easy user interface represented chiefly by one dialog box, displayed in Figure 1-2.

The upper part of the dialog box contains a short version of the instructions provided here. Once data are on the Clipboard, check the WCanoImp options that are appropriate for your situation. The first option (*Each column is a Sample*) applies only if you have your matrix transposed with respect to the form described above. This might be useful if you do not have many samples (because Microsoft Excel, for example, limits the number of columns to 256) but a high number of variables. If you do not have names of samples in the first column, you must check the second checkbox (i.e. ask to *Generate labels for: . . . Samples*), similarly check the third checkbox if the first row in the selected spreadsheet rectangle corresponds to the values in the first sample, not to the names of the variables. The last checkbox (*Save in Condensed Format*) governs the actual format used when creating the data file. The default format (used if this option is not checked) is the so-called full format; the alternative format is the condensed format. Unless you are worried about using too much hard disc space, it does not matter what you select here (the results of the statistical methods will be identical, whatever format is chosen).

After you have made sure the selected options are correct, you can proceed by clicking the *Save* button. You must first specify the name of the file to be generated and the place (disc letter and folder) where it will be stored. WCanoImp then requests a simple description (one line of ASCII text) for the data set being generated. This one line then appears in the analysis output and reminds you of the kind of data being used. A default text is suggested in case you do not care about this feature. WCanoImp then writes the file and informs you about its successful creation with another dialog box.

1.8. Transformation of species data

As will be shown in Chapter 3, ordination methods find the axes representing regression predictors that are optimal for predicting the values of the response variables, i.e. the values in the species data. Therefore, the problem of selecting a transformation for the response variables is rather similar to the problem one would have to solve if using any of the species as a single response variable in the (multiple) regression method. The one additional restriction is the need to specify an identical data transformation for all the response variables ('species'), because such variables are often measured on the same scale. In the unimodal (weighted averaging) ordination methods (see Section 3.2), 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. Therefore, CANOCO provides a flexible log-transformation formula:

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

You should specify the values of A and C so that after the transformation is applied to your data values (y), the result (y') is always greater or equal to zero. The default values of both A and C are 1.0, which neatly map the zero values again to zero, and other values are 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 10.0. But 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.

The question of when to apply a log transformation and when to use the original scale is not an easy one to answer and there are almost as many answers as there are statisticians. We advise you not to think so much about distributional properties, at least not in the sense of comparing frequency histograms of the variables with the 'ideal' Gaussian (Normal) distribution. Rather try to work out whether to stay on the original scale or to log-transform by using the semantics of the hypothesis you are trying to address.

As stated above, ordination methods can be viewed as an extension of multiple regression methods, so this approach will be explained in the simpler regression context. You might try to predict the abundance of a particular species in samples 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, your answer 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 $Y = \mathrm{B}_0 + \mathrm{B} \cdot X + E$. But in other cases, you might prefer to see the answer in a different form, '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'. Here 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**.

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 often quantified in the spreadsheets using corresponding ordinal levels (from 1 to 7 in this case). Note that this coding 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, 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 available in CANOCO is the square-root transformation. This might be the best transformation to apply to count data (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.

The console version of CANOCO 4.x also provides the rather general 'linear piecewise transformation' which allows you to approximate the more complicated transformation functions using a poly-line with defined coordinates of the 'knots'. This general transformation is not present in the Windows version of CANOCO, however.

Additionally, if you need any kind of transformation that is not provided by the CANOCO software, you might do it in your spreadsheet software and export the transformed data into CANOCO format. This is particularly useful in cases where your 'species data' do not describe community composition but something like chemical and physical soil properties. In such a case, the variables have different units of measurement and different transformations might be appropriate for different variables.

1.9. Transformation of explanatory variables

Because the explanatory variables ('environmental variables' and 'covariables' in CANOCO terminology) are assumed not to have a uniform scale, you need to select an appropriate transformation (including the popular 'no transformation' choice) individually for each such variable. CANOCO does not provide this feature; therefore, any transformations on the explanatory variables must be done before the data are exported into a CANOCO-compatible data file.

But you should be aware that after CANOCO reads in the environmental variables and/or covariables, it centres and standardizes them all, to bring their means to zero and their variances to one (this procedure is often called *standardization to unit variance*).