Multivariate analysis

Dr. Daniel Borcard
Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre Ville
Montréal QC
H3C 3J7 Canada
daniel.borcard@umontreal.ca

Foreword: this document is heavily based on the following book, with permission of Pierre Legendre:

Legendre, P. & L. Legendre. 1998. Numerical ecology. Second English Edition. Elsevier, Amsterdam.

This book is a MUST! It contains, among many other topics, all the mathematical developments that have been deliberately excluded from this summary. Many of the paragraphs, phrases, and several figures and tables come directly from this book. To Pierre Legendre I express my deepest thanks for his permission to use this material, as well as for his willingness to answer many, sometimes contorted questions.

i. Additional references, software and definitions

i.1 Additional references

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

Mainly regression, ordination and spatial analysis.

Legendre, L. & P. Legendre. 1984. Ecologie numérique. Vol. 1 & 2. Masson, Paris.

The French edition, still useful; many important topics were not availabe at that time, though.

Ter Braak, C.J.F. & P. Smilauer. 2002. CANOCO reference manual and CanoDraw for Windows user's guide: software for canonical community ordination (version 4.5). Microcomputer Power, Ithaca.

Much more than a simple user's manual of the latest version of the time-honored program Canoco. Very important for people interested in canonical analysis of experimental data coming from ANOVA designs.

i.2 Software

• **Excel** XP for Windows (or for Mac OSX)

Preparation of tabular data, simple statistics and graphics.

• **R** 2.6.0 for Windows, Mac OS X or Linux

General statistics, matrix algebra, multivariate statistics (cluster analysis, ordination...). Open source, clone of S-Plus.

Packages dedicated to numerical ecology: vegan, labdsv, ade4…

http://cran.r-project.org/

http://cc.oulu.fi/~jarioksa/softhelp/vegan.html

http://biomserv.univ-lyon1.fr/~dray/software.php

http://ecology.msu.montana.edu/labdsv/R/

http://pbil.univ-lyon1.fr/R/enseignement.html

• **CANOCO** 4.5 for Windows (3.1 for Mac OS)

Constrained and unconstrained ordination

Commercial software developed by Cajo Ter Braak

http://www.plant.dlo.nl/default.asp?section=products&page=/products/canoco/right.htm, http://www.microcomputerpower.com

• **R Package** 4.0d8 for Mac OS (Classic environnement)

Assocation matrices (many coefficients), constrained and unconstrained clustering, unconstrained ordination, spatial analysis, ordination graphical support, Mantel test. Freeware, work in progress (P. Legendre and Ph. Casgrain). Not to be confused with the R language!

http://www.bio.umontreal.ca/legendre/

i.3 Definitions

Numerical ecology: "the field of quantitative ecology devoted to the numerical analysis of ecological data sets. (...) The purpose of numerical ecology is to describe and interpret the structure of data sets by combining a variety of numerical approaches. Numerical ecology differs from descriptive or inferential *biological statistics* in that it extensively uses non-statistical procedures, and systematically combines relevant multidimensional statistical methods with non-statistical numerical techniques (e.g. cluster analysis) (...) " (Legendre & Legendre, 1998).

Let us add that a great number of the methods used in numerical ecology, especially the new approaches developed since the 80's, have been devised by ecologists (and not pure statisticians), in response to specific ecological problems.

Multivariate, multidimensional analysis: methods of numerical analysis addressing whole data tables where every observation, i.e. every sampling or experimental unit is characterised by several variables: species abundances, climatic measures, and so on.

1. The data

1.1 Data matrices

Instead of treating dependent variables one at a time, multivariate analysis considers **data tables**. The ecological data table is generally a **rectangular matrix** of the following form (Table I):

Table I - Structure of an ecological data table

	Descriptors				
Objects	Variable 1	Variable 2	Variable *j*		Variable *p*
Object 1	y_{11}	y_{12}	... y_{1j}	...	y_{1p}
Object 2	y_{21}	y_{22}	... y_{2j}	...	y_{2p}
.					
Object *i*	y_{i1}	y_{i2}	... y_{ij}	...	y_{ip}
.					
Object *n*	y_{n1}	y_{n2}	... y_{nj}	...	y_{np}

The **objects** are the observations (sites, relevés...).

The best-known example of an ecological data table is the one where the variables are **species** (represented as counts, presence-absence, or any appropriate form of numerical coding) and the objects are sites, vegetation relevés, field observations, traps, and so on.

An ecological data table can also be made of **environmental variables** (climate, chemical variables...) that will be used either to explain the structure of a species table, or directly to characterise a group of sites.

Finally, another such table may contain the **geographical coordinates** or any appropriate coding of the spatial structure of the data set.

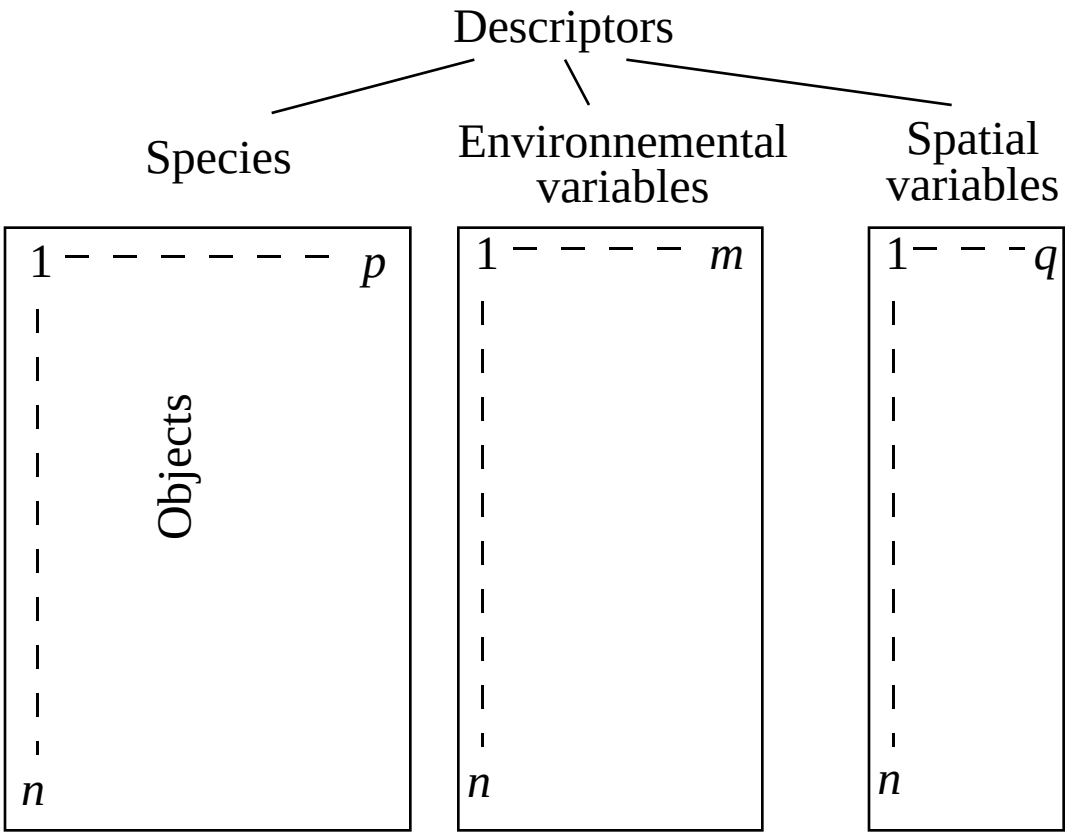

Figure 1 - The ecologist's data matrices.

The methods addressed in this document are aimed at the following goals:

- measurement of **resemblance** among objects or variables of a data table;

- **clustering** of the objects or variables according to these resemblances;

- **ordination** in a reduced space allowing to reveal their main structures (especially gradients);

- **modelling** of the relationships between response data tables and explanatory variables;

- **tests** of these relationships for statistical significance.

1.2 Data transformation

There are instances where one needs to transform the data prior to analysis. The main reasons are given below.

1. Make comparable descriptors that have been measured in different units

This is often done using **ranging** or **standardization** of the variables. It is useful because many methods are sensitive to the scale of measurements of the variables. While this is sometimes a desirable property, in other cases one prefers to assess the ecological structures independently of the units of the variables.

Ranging is made of two operations: a) **translation**: subtract its minimum from each variable; b) **expansion**: divide the values by the variable's range. This constrains the values of the variable to the interval [0;1]:

$$y_i' = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \quad \text{or} \quad y_i' = \frac{y_i}{y_{\max}}$$

The left-hand transformation above is used on variables where the zero value is chosen arbitrarily (called *interval scale* variables; an example is the Celsius temperature scale). For variables with a true zero and no negative values (called *relative scale* variables), ranging can often be simplified as in the right-hand equation above.

Standardization: subtract the mean of the variable from each value (i.e. **centre** the variable), and divide the results by the standard deviation of the variable (i.e. **scale** the variable). This yields the so-called "z-scores":

$$y_i' = z_i = \frac{y_i - \bar{y}}{s_y}$$

This results in a variable that has zero mean and unit variance (and hence standard deviation = 1 as well). Therefore, all variables that have been standardized can be directly compared and used together in methods that are sensitive to differences in scales of measurement, since they are now dimensionless and expressed as standard deviation units.

2. Normalize the data and stabilize their variance

This is done to make the frequency distribution of the values look like a normal curve - or, at least, as symmetric as possible. This is done because several multivariate methods used in ecology have been developed under the assumption that the variables are normally distributed. Full normality is generally not necessary, however, but the performance of these methods is better with unskewed data.

Normalizing can be done in different ways, that require the examination of the frequency distribution of the data. In many cases, ecologists encounter data that are strongly skewed to the right (long tail in the high values), because, in a sample set of species abundances, a species is abundant in a few observation units, fairly abundant in more, present in even more, and absent from many units. Depending on the skewness observed and the types of data, various correcting transformations can be applied.

• **Square root transformation** (Figure 2): $y'_i = \sqrt{(y_i+c)}$. The least drastic transformation, used when the data have a Poisson distribution; the constant c must be added to the data if there are negative values. So one first makes a *translation* of the data (c being equal to the absolute value of the most negative observation) prior to the transformation itself.

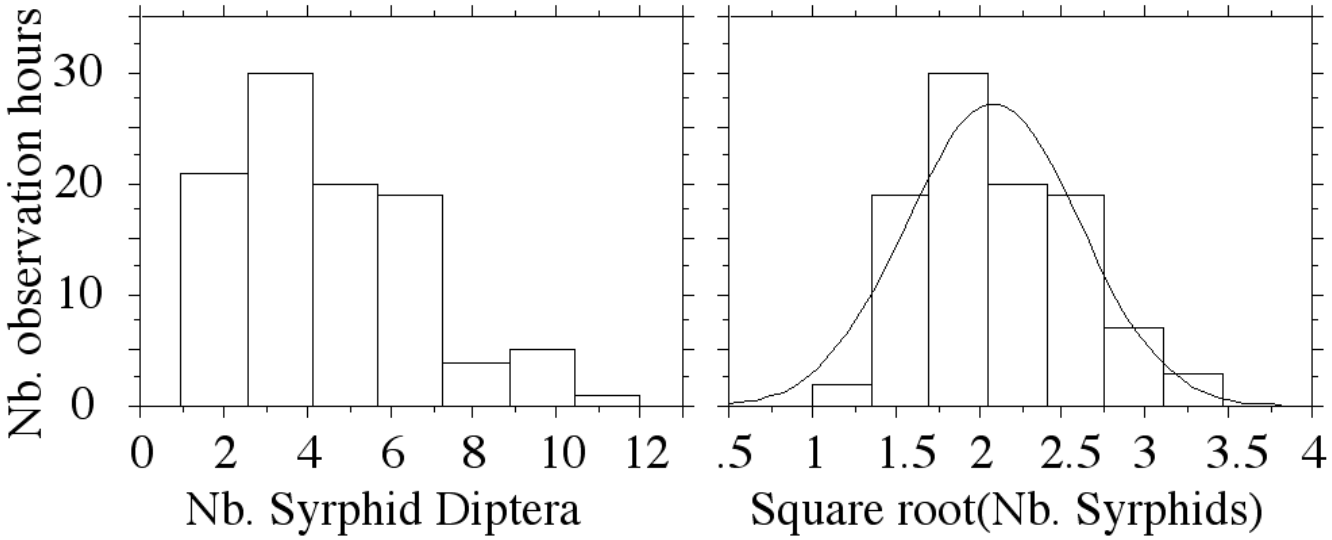

Figure 2 - The square root transformation.

• **Log transformation**: $y'_i = \ln(y_i+c)$. Frequently applied to species abundance data, of which many tend to follow a lognormal distribution. The base of the logarithm is not important, but the mostly used are the Napierian (natural) logarithms. The constant c is added to make the data strictly positive. With species abundance data, c is generally set equal to 1. Thus, zero values are translated to 1, and become zero again with the log transformation.

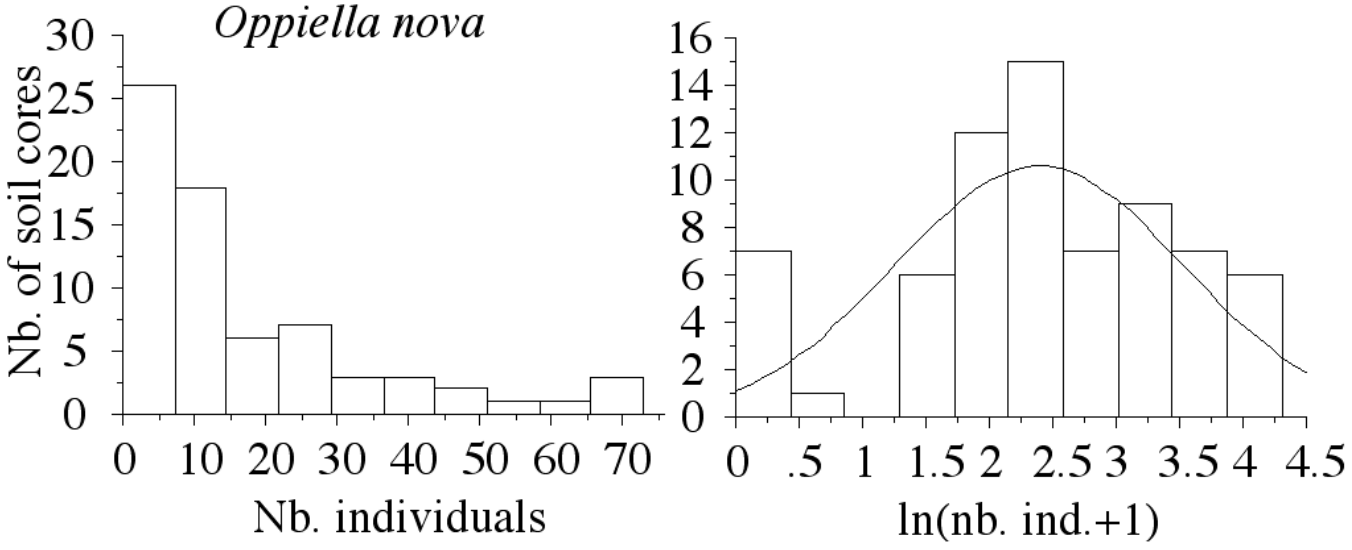

Figure 3 - The log transformation.

• **Arcsine transformation**: appropriate for percentages or proportions (which are generally platykurtic), but the analytical results based on arcsine-transformed data may be difficult to interpret: $y_i' = \arcsin\sqrt{y_i}$.

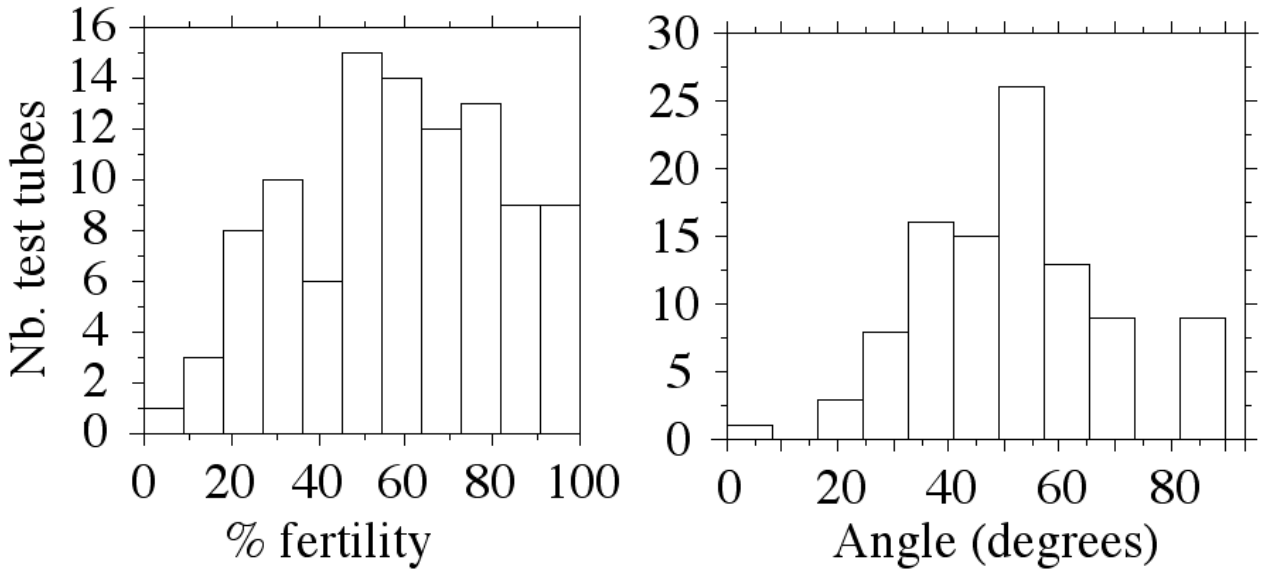

Figure 4 - The arcsine transformation. Data: Sokal & Rohlf (1981).

• **Box-Cox transformation**: when there is no *a priori* reason to select one of the transformations above, the Box-Cox method allows one to empirically (and iteratively) estimate the most appropriate exponent of the following general transformation function:

$$y_i' = (y_i^{\gamma} - 1)/\gamma \quad (\text{for } \gamma \neq 0)$$

$$y_i' = \ln(y_i) \quad (\text{for } \gamma = 0)$$

Normalizing transformation generally also have the property of *stabilizing the variances*; homoscedasticity (stability or homogeneity of variances) is an essential property of the data for several analysis, including ANOVA and its multivariate counterparts, and this, even if the tests are conducted using permutations (see Chapter 5).

3. Linearize the relationships among variables

Comparison coefficients like covariance or Pearson correlation are made to detect linear relationships. Thus, if the relationships among variables are monotonic but nonlinear, a transformation may be applied. For instance, if a response variable is an exponential function of an independent variable, then the response variable may be log-transformed. The reverse may occur also. Note that it will be easier to interpret the results if the transformation applied has a ground in ecological theory. An example is the Malthusian exponential growth curve:

$N_t = N_0 e^{rt}$

Data of a time series showing this curve may be log-transformed so that $\ln(N_t)$ becomes linearly related to time t: $\ln(N_t) = \ln(N_0) + rt$.

4. Modify the weights of the variables or objects

Standardization, log transformation or exponential transformation also have the effect of modifying the relative weight of the variables. Other transformations may also explicitly change the weight of the observations, as for instance the **normalization** of the object or variable vectors to 1 (do not confuse with the normalizing transformations above!). This operation consists in dividing each value of the vector by the vector's length (called the **norm** of the vector), which is defined following Pythagora's formula:

$$\text{Vector norm} = \sqrt{b_1^2 + b_2^2 + ... + b_n^2}$$

where b are the observations and 1, 2... are the object indices (so this example deals with a variable).

The normalized vector is thus defined as:

$$\begin{bmatrix} b_1/\sqrt{b_1^2+b_2^2+\ldots+b_n^2} \\ b_2/\sqrt{b_1^2+b_2^2+\ldots+b_n^2} \\ \ldots \\ b_n/\sqrt{b_1^2+b_2^2+\ldots+b_n^2} \end{bmatrix} = \frac{1}{\sqrt{b_1^2+b_2^2+\ldots+b_n^2}} \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_n \end{bmatrix}$$

The length of any normalized vector, in the *n*-dimensional space, is 1.

5. Recode semi-quantitative variables as quantitative

In many instances variables are measured on a semi-quantitative (ordinal) scale, generally because the added precision of a quantitative measurement would not justify the additional cost or difficulty to gather it. Such ordinal measurements are often devised in such away that the intervals between the classes follow a known distribution (for instance a variable of abundance classes going from 0 for "absent" to to 5 for "very abundant" may follow a logarithmic transformation of the real abundances). In such cases a back-transformation is possible, but one has to be conscious that this operation does not restore a precision that the original measurements did not have in the first place!

An complex example is the transformation of Braun-Blanquet's phyto-sociological scale into quantitative values (Van Der Maarel 1977):

Skala Braun-Bl.	Deck-ung, %	Ordinal-skala x	Transformation $y = x^w$					
			w=0	w=0.25	w=0.5	w=1	w=2	w=4
leer	0.0	0	0.0	0.00	0.00	0	0	0
r	()	1	1.0	1.00	1.00	1	1	1
+	0.1	2	1.0	1.19	1.41	2	4	16
1	5.0	3	1.0	1.32	1.73	3	9	81
2m		4	1.0	1.41	2.00	4	16	256
2 2a	17.5	5	1.0	1.50	2.24	5	25	625
2b		6	1.0	1.57	2.45	6	36	1296
3	37.5	7	1.0	1.63	2.65	7	49	2401
4	62.5	8	1.0	1.68	2.83	8	64	4096
5	87.5	9	1.0	1.73	3.00	9	81	6561

Table II - Transformation of Braun-Blanquet's scores into quantitative scores (preceding page).

6. Binary coding of nominal variables

Many statistical packages incorrectly interpret or do not accept multistate nominal variables (see Section 2.2) whose classes are coded as incremental numbers or as chains of characters. If it is not possible to declare nominal variables as such (i.e. "factor"-type variables), one must recode these variables into a series of dummy binary variables:

Table III: binary coding of a nominal variable. Note that here 3 dummy variables are sufficient, the fourth one being collinear to the others. The fourth one is often discarded by computer programs, or the analysis can simply not be run with it.

One nominal variable		4 dummy binary variables			
Modality	Code	Calcosol	Brunisol	Neoluvisol	*Calcisol*
Calcosol	1	1	0	0	*0*
Brunisol	2	0	1	0	*0*
Neoluvisol	3	0	0	1	*0*
Calcisol	4	0	0	0	*1*