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 2 Offprint

The Four Corners of Multivariate Analysis

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

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

Fundación **BBVA**

Chapter 2

The Four Corners of Multivariate Analysis

Multivariate analysis is a wide and diverse field in modern statistics. In this chapter we shall give an overview of all the multivariate methods encountered in ecology. Most textbooks merely list the methods, whereas our approach is to structure the whole area in terms of the principal objective of the methods, divided into two main types – functional methods and structural methods. Each of these types is subdivided into two types again, depending on whether the variable or variables of main interest are continuous or categorical. This gives four basic classes of methods, which we call the "four corners" of multivariate analysis, and all multivariate methods can be classified into one of these corners. Some methodologies, especially more recently developed ones that are formulated more generally, are of a hybrid nature in that they lie in two or more corners of this general scheme.

Contents

The basic data structure: a rectangular data matrix

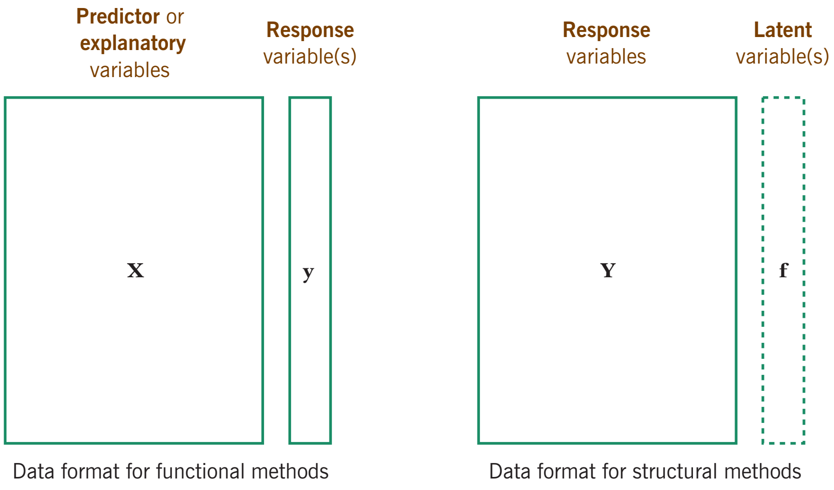
In multivariate statistics the basic structure of the data is in the form of a cases-by-variables rectangular table. This is also the usual way that data are physically stored in a computer file, be it a text file or a spreadsheet, with cases as rows and variables as columns. In some particular contexts there are very many more variables than cases and for practical purposes the variables are defined as rows of the matrix: in genetics, for example, there can be thousands of genes observed on a few samples, and in community ecology species (in their hundreds) can be listed as rows and the samples (less than a hundred) as columns of the data table. By convention, however, we will always assume that the rows are the cases or sampling units of the study (for example, sampling locations, individual animals or plants, laboratory samples), while the columns are the variables (for example: species, chemical compounds, environmental parameters, morphometric measurements).

Variables in a data matrix can be on the same measurement scale or not (in the next chapter we treat measurement scales in more detail). For example, the matrix might consist entirely of species abundances, in which case we say that we have "same-scale" data: all data in this case are counts. A matrix of morphometric measurements, all in millimetres, is also same-scale. Often we have a data matrix with variables on two or more different measurement scales – the data are then called "mixed-scale". For example, on a sample of fish we might have the composition of each one's stomach contents, expressed as a set of percentages, along with morphometric measurements in millimetres and categorical classifications such as sex (male or female) and habitat (littoral or pelagic).

Functional and structural methods

We distinguish two main classes of data matrix, shown schematically in Exhibit 2.1. On the left is a data matrix where one of the observed variables is separated from the rest because it has a special role in the study – it is often called a *response variable*. By convention we denote the response data by the column vector **y**, while data on the other variables – called *predictor*, or *explanatory, variables* – are gathered in a matrix **X**. We could have several response variables, gathered in a matrix **Y**. On the right is a different situation, a data matrix **Y** of several response variables to be studied together, with no set of explanatory variables. For this case we have indicated by a dashed box the existence of an unobserved variable

Exhibit 2.1:
Schematic diagram of the two main types of situations in multivariate analysis: on the left, a data matrix where a variable **y** *is singled out as being a response variable and can be partially explained in terms of the variables in* **X**. *On the right, a data matrix* **Y** *with a set of response variables but no observed predictors, where* **Y** *is regarded as being explained by an unobserved, latent variable* **f**


Predictor or explanatory variables
Response variable(s)
X
y
Data format for functional methods
Response variables
Latent variable(s)
Y
f
Data format for structural methods

f, called a *latent variable*, which we assume has been responsible, at least partially, for generating the data **Y** that we have observed. The vector **f** could also consist of several variables and thus be a matrix **F** of latent variables.

We call the multivariate methods which treat the left hand situation *functional methods*, because they aim to come up with a model which relates the response variable **y** as a function of the explanatory variables **X**. As we shall see in the course of this book, the nature of this model need not be a mathematical formula, often referred to as a *parametric* model, but could also be a more general *nonparametric* concept such as a tree or a set of smooth functions (these terms will be explained more fully later). The methods which treat the right hand situation of Exhibit 2.1 are called *structural methods*, because they look for a structure underlying the data matrix **Y**. This latent structure can be of different forms, for example gradients or typologies.

One major distinction within each of the classes of functional and structural methods will be whether the response variable (for functional methods) or the latent variable (for structural methods) is of a continuous or a categorical nature. This leads us to a subdivision within each class, and thus to what we call the "four corners" of multivariate analysis.

The four corners of multivariate analysis

Exhibit 2.2 shows a vertical division between functional and structural methods and a horizontal division between continuous and discrete variables of interest, where "of interest" refers to the response variable(s) **y** (or **Y**) for functional methods and the latent variable(s) **f** (or **F**) for the structural methods (see Exhibit 2.1). The four quadrants of this scheme contain classes of methods, which we shall treat one at a time, starting from regression at top left and then moving clockwise.

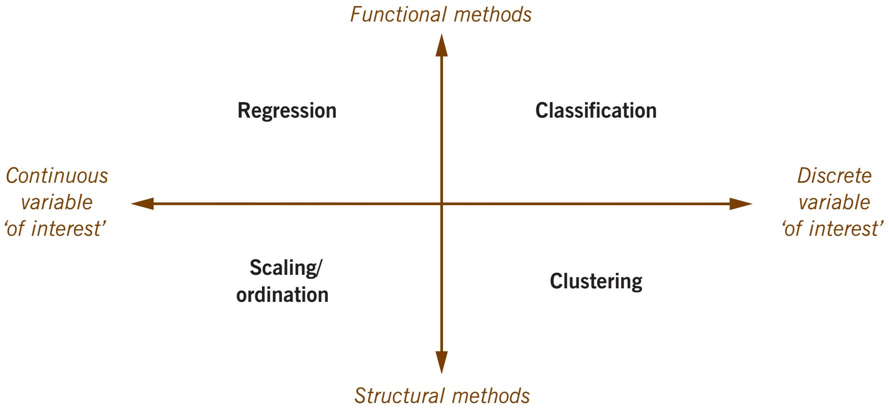


Exhibit 2.2:
The four corners of multivariate analysis. Vertically, functional and structural methods are distinguished. Horizontally, continuous and discrete variables of interest are contrasted: the response variable(s) in the case of functional methods, and the latent variable(s) in the case of structural methods

Regression: Functional methods explaining a given continuous variable

At top left we have what is probably the widest and most prolific area of statistics, generically called *regression*. In fact, some practitioners operate almost entirely in this area and hardly move beyond it. This class of methods attempts to use multivariate data to explain one or more continuous response variables. In our context, an example of a response variable might be the abundance of a particular plant species, in which case the explanatory variables could be environmental characteristics such as soil texture, pH level, altitude and whether the sampled area is in direct sunshine or not. Notice that the explanatory variables can be of any type, continuous or categorical, whereas it is the continuous nature of the response variable – in this example, species abundance – that implies that the appropriate methodology is of a regression nature.

In this class of regression methods are included multiple linear regression, analysis of variance, the general linear model and regression trees. Regression methods can have either or both of the following purposes: to explain relationships and/or to predict the response from new observations of the explanatory variables. For example, on the one hand, a botanist can use regression to study and quantify the relationship between a plant species and several variables that are believed to influence the abundance of the plant. But on the other hand, the objective could also be to ask “what if?” type questions: what if the rainfall decreased by so much % and the pH level rose to such-and-such a value, what would the average abundance of the species be, and how accurate is this prediction?

Classification: Functional methods explaining a given categorical variable

Moving to the top right corner of Exhibit 2.2, we have the class of methods analogous to regression but with the crucial difference that the response variable is not continuous but categorical. That is, we are attempting to model and predict a variable that takes on a small number of discrete values, not necessarily in any order. This area of *classification* methodology is omnipresent in the biological and life sciences: given a set of observations measured on a person just admitted to hospital having experienced a heart attack – age, body mass index, pulse, blood pressure and glucose level, having diabetes or not, etc. – can we predict whether the patient will survive in the next 24 hours? Having found a fossil skull at an archeological site, and made several morphometric measurements, can we say whether it is a human skull or not, and with what certainty?

The above questions are all phrased in terms of predicting categories, but our investigation would also include trying to understand the relationship between a categorical variable, with categories such as “healthy” and “sick”, and a host of variables that are measured on each individual. Especially, we would like to know which of these variables is the most important for discriminating between the categories. Classification methods can also be used just to quantify differences between groups, as in Exhibits 1.5 and 1.8 of Chapter 1. There we observed some differences between the sediment types for one variable at a time; the multivariate challenge will be to see if we can quantify *combinations* of variables that explain group differences.

Clustering: Structural methods uncovering a latent categorical variable

We now move down to the structural methods, where the unobserved latent variable **f** is sought that "explains" the many observed variables **Y**. This is a much more subtle area than the functional one, almost abstract in nature: how can a variable be "unobserved"? Well, let us suppose that we have collected a whole bunch of data on clams found in the Arctic. Are they all of the same species? Suppose they are not and there are really two species involved, but we cannot observe for a particular clam whether it is of species A or B, we just do not know. So species is an unobserved categorical variable. Because it is categorical, we are in the bottom right area of the scheme of Exhibit 2.2. The idea of clustering is to look for similarities between the individual clams, not on single variables but across all measured variables. Can we come up with a grouping (i.e., *clustering*) of the clams into two clusters, each of which consists internally of quite similar individuals, but which are quite different if we compare individuals from different clusters? This is the objective of cluster analysis, to create a categorical structure on the data which assigns each individual to a cluster category. Supposing that the cluster analysis does come up with two clusters of clams, it is then up to the marine biologist to consider the differences between the two clusters to assess if these are large enough to warrant evidence of two different species.

Cluster analysis is often called *unsupervised learning* because the agenda is open to whether groups really do exist and how many there are; hence we are learning without guidance, as it were. Classification methods, on the other hand, are sometimes called *supervised learning* because we know exactly what the groups are and we are trying to learn how to predict them.

Scaling/ordination: Structural methods uncovering a latent continuous variable

The final class of methods, at bottom left in Exhibit 2.2, comprise the various techniques of *scaling*, more often referred to as *ordination* by ecologists. Ordination is just like clustering except the structures that we are looking for in the data are not of a categorical nature, but continuous. Examples of ordination abound in environmental science, so this will be one of the golden threads throughout this book. The origins of scaling, however, come from psychology where measurement is an issue more than in any other scientific discipline. It is relatively simple for a marine biologist to measure a general level of "pollution" – although the various chemical analyses may be expensive, reliable figures can be obtained of heavy metal concentrations and organic materials in any given sample. A psychologist interested in emotions such as anxiety or satisfaction, has a much more difficult job arriving at a reliable quantification. Dozens of measurements could be made to assess the level of anxiety, for example, most of them "soft" in the sense that they could be answers to a battery of questions on how the respondent feels. Scaling attempts to discern whether there is some underlying dimension (i.e., a scale) in these data which is ordering the respondents from one extreme to another. If this dimension can be picked up, it is then up to the psychologist, to decide whether it validly orders people along a continuous construct that one might call “anxiety”.

In the large data sets collected by environmental biologists, the search for continuous constructs can be the identification of various environmental *gradients* in the data, for example pollution or temperature gradients, and of geographical gradients (e.g., north–south). Almost always, several gradients (i.e., several continuous latent variables **f**) can be identified in the data, and these provide new ways of interpreting the data, not in terms of their original variables, which are many, but in terms of these fewer latent dimensions. Because of the importance of ordination and reduction of dimensionality in environmental research, a large part of this book will be devoted to this area.

Hybrid methods

It is in the nature of scientific endeavour that generalizations are made that move the field ahead while including everything that has been discovered before. Statistics is no exception and there are many examples of methods developed as generalizations of previous work, or a gathering together of interconnected methodologies. “General linear modelling” and “generalized linear modelling” (similar in name but different in scope) are two such examples.

In classical linear regression of a continuous response variable there are several variants: multiple regression (where all the explanatory variables are continuous), analysis of variance (ANOVA, where all the explanatory variables are categorical), and analysis of covariance (ANCOVA, where the explanatory variables are continuous and categorical). Each of these has its own quirks and diagnostics and terminology. Design and analysis of experiments usually involve ANOVA or ANCOVA, where the cases are assigned to various treatments in order to be able to estimate their effects. All of these methods are subsumed under the umbrella of the *general linear model*, which falls into the regression corner of Exhibit 2.2.

A more fundamental gathering together of methodologies has taken place in the form of *generalized* linear modelling. Many techniques of regression and classification, which we grouped together as functional methods, have an inherent similarity in that the explanatory variables are combined linearly in order to make models or predictions of the response variable. The aspect that distinguishes them is how that linear function is used to connect with the response variable, and what probability distribution is assumed for the conditional distributions of the response. The *generalized linear model* involves firstly the choice of a function that acts as the link between the mean of the response variable and the predictors, and secondly the choice of a distribution of the response around this mean – these choices lead to many well-known modelling methods as special cases. Generalized linear modelling straddles both functional corners of our four corner multivariate analysis scheme. Multiple linear regression is the simplest generalized linear model, while logistic regression (when responses are categorical) and Poisson regression (when responses are counts) are other examples – more details will be given in Chapter 18.

Well-known methods in environmental science are *canonical correspondence analysis* and *redundancy analysis* (see Chapter 15). These are basically ordination methods but force the ordination scales to be functions of observed explanatory variables, which recalls the idea of regression. Hence canonical correspondence analysis can be said to straddle the upper and lower corners on the left of our scheme. The method of *partial least squares* has a similar objective, but allows the inclusion of a very large number of explanatory variables.

Finally, the generalization of all generalizations is potentially *structural equation modelling*. We say "potentially" because it is presently covering at least the two left hand continuous corners of our scheme and, as the field develops, moving to cover the right hand ones as well. This area of methodology and its accompanying terminology are very specific to psychological and sociological research at the moment, but could easily find wider use in the environmental sciences as more options are added to handle count and categorical response data.

SUMMARY: The four corners of multivariate analysis

1. Methods of multivariate analysis treat rectangular data matrices where the rows are usually cases, individuals, sampling or experimental units, and the columns are variables.

2. A basic classification of methods can be achieved by first distinguishing the overall objective of a study as either (i) explaining an observed "response" variable in terms of the others, or (ii) ascertaining the inherent structure in the form of a "latent" variable that underlies the set of observed variables. This separates functional from structural methods, respectively.

3. Functional and structural methods can be subdivided into those where (in the case of functional methods) the response variable is continuous or categorical, or (in the case of structural methods) where the identified latent structure is of a continuous or categorical nature.

4. Thus four main classes of methods exist: functional methods explaining a continuous variable (regression and related methods), functional methods explaining a categorical variable (classification), structural methods with latent

structure that is continuous (scaling/ordination) and structural methods with latent categorical structure (clustering).

5. Several general methodologies, such as general and generalized linear models, canonical correspondence analysis, partial least squares and structural equation modelling, can cover more than one of these classes.

List of Exhibits