MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 1032–1057 doi: 10.1111/mec.13536

INVITED REVIEWS AND SYNTHESES

Application of multivariate statistical techniques in microbial ecology

O. PALIY and V. SHANKAR

Department of Biochemistry and Molecular Biology, Boonshoft School of Medicine, Wright State University, 260 Diggs Laboratory, 3640 Col. Glenn Hwy, Dayton, OH 45435, USA

Abstract

Recent advances in high-throughput methods of molecular analyses have led to an explosion of studies generating large-scale ecological data sets. In particular, noticeable effect has been attained in the field of microbial ecology, where new experimental approaches provided in-depth assessments of the composition, functions and dynamic changes of complex microbial communities. Because even a single high-throughput experiment produces large amount of data, powerful statistical techniques of multivariate analysis are well suited to analyse and interpret these data sets. Many different multivariate techniques are available, and often it is not clear which method should be applied to a particular data set. In this review, we describe and compare the most widely used multivariate statistical techniques including exploratory, interpretive and discriminatory procedures. We consider several important limitations and assumptions of these methods, and we present examples of how these approaches have been utilized in recent studies to provide insight into the ecology of the microbial world. Finally, we offer suggestions for the selection of appropriate methods based on the research question and data set structure.

Keywords: microbial communities, microbial ecology, microbiota, multivariate, ordination, statistics

Received 1 May 2015; revision received 15 December 2015; accepted 22 December 2015

Introduction

The past decade has seen significant progress in ecological research due in part to the advent and increased utilization of novel high-throughput experimental technologies. With approaches such as high-throughput ‘next-generation’ sequencing, oligonucleotide and DNA microarrays, high sensitivity mass spectrometry, and nuclear magnetic resonance analysis, researchers are able to generate massive amounts of molecular data even in a single experiment. These methods have created an especially powerful effect on the field of microbial ecology, where we now can apply DNA, RNA, protein, and metabolite identification and measurement techniques to the whole microbial community without a need to separate or isolate individual community members. These technologies provided foundation to many groundbreaking advances in our understanding of microbial community organization, function, and interactions within a community and with other organisms. Examples include the assessment of marine microbiota response to the Deepwater Horizon oil spill (Mason *et al.* 2014), functional analysis of microbiomes in different soils (Fierer *et al.* 2012), identification of enterotypes in the human intestinal microbiota (Arumugam *et al.* 2011), detection of seasonal fluctuations in oceanic bacterioplankton (Gilbert *et al.* 2012) and discovery of the loss of gut microbial interactions in human gastrointestinal diseases (Shankar *et al.* 2013, 2015).

Because large amount of data is generated even in a single high-throughput experiment, powerful statistical tools are needed to examine and interpret the results. Because many different variables such as species, genes, proteins or metabolites are measured in each sample or site, the analysis of these data sets is generally performed using multivariate statistics

Correspondence: Oleg Paliy, Fax: (1) 937 775 3730;
E-mail: oleg.paliy@wright.edu

Box 1. Terminology used in multivariate statistical analyses

- Biplot—a two-dimensional diagram of the ordination analysis output that simultaneously shows variable positioning and object positioning in a reduced dimensionality space.
- Canonical analysis—a general term for statistical technique that aims to find relationship(s) between sets of variables by searching for latent (hidden) gradients that associate these sets of variables.
- Constrained and unconstrained ordination—the constrained multivariate techniques attempt to 'explain' the variation in a set of response variables (e.g. species abundance) by the variation in a set of explanatory variables (e.g. environmental parameters) measured in the same set of objects (e.g. samples or sites). The matrix of explanatory variables is said to 'constrain' the multivariate analysis of the data set of response variables, and the output of constrained analysis typically displays only the variation that can be explained by constraining variables. In contrast, the unconstrained multivariate techniques only examine the data set of response variables, and the output of unconstrained analysis reflects overall variance in the data.
- Gradient analysis—this term describes the study of distribution of variable values in the data set along gradients. As the goal of ordination analysis is to order objects along the main gradients of dispersion in the data set, both of these terms can be used synonymously. Two different types of gradient analysis are usually recognized:
  - Indirect gradient analysis utilizes only one data set of measured variables. This term is synonymous with 'unconstrained ordination'.
  - Direct gradient analysis in contrast uses additionally available data to guide (direct) the analysis of the data set of measured variables. It produces axes that are *constrained* to be a function of explanatory variables. This term is synonymous with 'constrained ordination'.
- Data transformation—this term describes the process of applying a mathematical function to the full set of measured values in a systematic way. See Box 2 for a description of common data transformations.
- Distance—it quantifies the dissimilarity between objects in a specific coordinate system. Objects that are similar have small in-between distance; objects that are different have large distance between them. For normalized distances, Distance (X,Y) can be related to Similarity (X,Y) as either $D = 1 - S$, $D = \sqrt{(}1 - S)$ or $D = \sqrt{(}1 - S^2)$. Many different mathematical functions can be used to calculate distances among objects or variables (Legendre & Legendre 2012). The choice of distance measure has a profound effect on the output of multivariate analysis and should be chosen based on the characteristics of the studied ecological data set.
- Eigenvector and eigenvalue—in ordination methods such as PCA, eigenvectors represent the gradients of data set dispersion in ordination space and are used as ordination axes, and eigenvalues designate the 'strength' of each gradient.
- Explanatory/predictor/independent variable—all these terms describe the type of variables that explain (predict) other variables.
- Ordination—a general term that can be described as 'arranging objects in order' (Goodall 1954). The goal of ordination analysis is to generate a reduced number of new synthetic axes that are used to display the distribution of objects along the main gradients in the data set.
- Orthogonal—mathematical term that means 'perpendicular' or 'at right angle to'. Due to this feature, the orthogonal variables are *linearly* independent (Rodgers *et al.* 1984).
- Randomization tests—a group of related tests of statistical significance that are based on the randomization of measured data values to assess whether the value of a calculated metric (such as species diversity) can be obtained by chance.
- Response/dependent variable—both of these terms describe the main measured variables in a study. Many multivariate analyses examine the relationship between these variables and other variables that are said to 'predict' or 'explain' these measured variables (such as environmental factors). Thus, the measured variables are 'dependent on' or 'responding to' the values of the explanatory/predictor variable(s).
- Triplot—a two-dimensional diagram of the multivariate analysis output that in addition to response variables and objects shown on biplot also displays explanatory variables.
- Unimodal distribution—a distribution with one peak on the variable density plot.
- Variance, variability and variation—the differences in the use of these three related terms can be described by the following statement: 'Variance is a statistical measure of data variation and dispersion, which describe the amount of variability in the data set'.

(definitions of most commonly used terms are provided in Box 1). Indeed, many types of standard multivariate statistical analyses have been employed for the assessment of such high-throughput data sets, and novel approaches are also being developed. The multitude of possible statistical choices makes it a daunting task for an investigator not experienced with these tools to pick a good technique to use. In this review, we provide short descriptions of the most frequently used multivariate statistical techniques, we compare different methods, and we present examples of how these approaches have been utilized in recent studies. This text is not meant to serve as an exhaustive overview of existing methods, as the landscape of currently available multivariate analyses is vast and growing. Rather, our goal is to familiarize the reader with the most commonly used approaches and to provide background on the differences among techniques and possible selection choices. While this review will primarily focus on the application of multivariate statistics to microbial ecological research, these techniques can be successfully applied to a wide variety of data sets in other ecological disciplines.

Types and properties of high-throughput ecological data sets

Although many different experimental tools can be utilized to obtain high-throughput ecological data, the output in most cases is presented as a matrix of positive numbers each representing either a measured value for variable i (e.g. mRNA or protein level, metabolite concentration or species abundance) in object j (e.g. sample or site), or a ratio of two measured x_i values between two objects. The former data set structure is common for outputs from high-throughput sequencing, mass spectrometry, NMR-based metabolomics and single-sample microarrays. The latter is usually obtained from two-colour microarrays or high-throughput quantitative PCR. The type and distribution of data should match

Box 2. Statistical assumptions and common data transformations

Generally, statistical approaches and tests can be divided into parametric and nonparametric. The parametric methods make an assumption that the data come from a population with a particular underlying probability distribution of the measured variable, and parametric statistics make inferences about the parameters of such distribution. This allows parametric tests to have higher statistical power and to make more accurate estimates. However, the measured variables have to meet the underlying assumptions for parametric methods to be properly applied. The most widespread assumptions are as follows:

- The population follows a defined distribution (this is often substituted by the expectation that sampled values follow a defined distribution). In most cases, either linear or unimodal distribution is expected.
- Variables are independent from each other.
- Population variances are equal or at least similar.
- Samples are drawn randomly from the population.

Unfortunately, biological data sets rarely conform to these assumptions. Most data sets are not normally distributed, variables that are displayed on the relative scale are no longer independent from each other, and often variances in different populations are not equal. In such cases, the original data set values can be transformed so that the scale and the distribution of transformed values conform better to the assumptions of a particular parametric statistic. Most common data transformations include the following:

- Log transformation $x'_i = \log_b(x_i + c)$—very useful for ratios; log bases 2, e or 10 are used most often; c is a small number added to deal with $x_i = 0$ cases.
- Root transformation $x'_i = (x_i)^{1/n}$—with $n \geq 1$, it compresses the spread of values in the right tail of the distribution.
- Power transformation $x'_i = (x_i)^n$—with $n \geq 1$, it has opposite to root transformation effect.
- Arcsine transformation $x'_i = \arcsin(x_i)$—is usually applied to per cent and proportion values.

These transformations are very useful for continuous variables, but are less suited for discrete response variables including count data (O'Hara & Kotze 2010). In addition, if the data set matrix contains many zero values (e.g. species missing from a particular habitat), many common methods of multivariate analysis such as PCA or RDA are not appropriate as they can create false distributions and outputs. For ecological data with many zeroes, Legendre & Gallagher (2001) recommend two special transformations that will allow these ordination techniques to be applied to the data set:

- Chord transformation: $x'_{ij} = x_{ij}/\sqrt{(\sum x^2_{ij})}$.
- Hellinger transformation $x'_{ij} = \sqrt{(x_{ij}/\sum x_{i+})}$; where i—species, j—object and i+ denotes all i's.

Alternatively, because nonparametric statistics are not based on parameterized probability distributions (they are 'distribution-free'), they can be directly applied to the original data set, albeit with a loss of statistical power (they are thus less likely to find a statistically significant difference). Examples of nonparametric methods are permutation tests and rank-based statistical analyses.

the assumptions of a particular statistical technique, and in some cases, data should be transformed prior to performing further tests and statistical analyses. We describe common statistical assumptions and few of the most widely used data transformation methods in Box 2.

Another important distinction among different data set structures is whether data points represent absolute or relative values. The former are usually raw signal values obtained through experimental measurements, while the latter are most frequently derived by scaling individual recorded signals to obtain the same total measured signal across objects. The need for signal transformation into a relative scale is often driven by the limitations of the experimental techniques. For example, multistep template preparation protocols used in high-throughput sequencing and microarray analyses can introduce an artificial bias into the observed sequence counts (Paliy & Foy 2011). Thus, we cannot compare raw signal values for the same variable across multiple samples and make comparative conclusions, because the observed higher or lower sequence abundance includes not only true DNA amount in the sample, but is also influenced by confounding variables such as sample DNA quality, amplification biases, variability in sequencing or hybridization efficiency and DNA concentration measurements. The effects of many such confounding factors can be removed if raw signal values for each object are converted into relative values through a simple $x_i' = x_i / \sum(x_i)$ transformation. While this transformation is very useful for simple across-object comparisons, it causes individual variables to lose their independency, which is one of the main assumptions in many statistical tests and analyses (Box 2). We describe the issues that can arise from the use of such compositional data, and approaches to mitigate these issues in Box 3.

Different types of multivariate approaches

Different available methods of large-scale data set analyses can be organized into groups based on various criteria such as technique goal (e.g. explore variance, interpret relationships, discriminate groups, test statistical significance), type of mathematical problem (regression, (partial) ordination, calibration, classification) or variable response model (e.g. linear, unimodal, mixture distribution) (ter Braak & Prentice 1988). Clear separation of the methods is hard to achieve, because the same technique can be used for several different purposes, methods can utilize different sets of parameters or interobject distance calculations, and many approaches are mathematically related. In this review, we distribute the described techniques into three categories based on the primary research objective of multivariate analysis.

1. Exploratory methods are used to explore the relationships among objects based on the values of variables measured in those objects. For example, soil samples (objects) collected in different landscapes can be compared based on the abundances of soil microbial species (variables) (Hartmann *et al.* 2014). These methods provide a useful visualization of object similarities, as similar objects are usually positioned close on the visualization plot, and dissimilar objects are apart from each other. Major gradients of data variation as well as object similarity can be evaluated. This category includes different unconstrained ordination techniques (Box 4) as well as cluster analyses (CLAs).
2. Interpretive methods are 'constrained' techniques, which in addition to the main set of measured variables, also use another set of additional explanatory variables (for example, known environmental gradients among objects) during the analysis. The aim of constrained ordination analyses is to find axes in the multidimensional data set space that maximize the association between the explanatory variable(s) and the measured variables (called response variables). Thus, the ordination axes are *constrained* to be functions of the explanatory variables. Coefficients for each explanatory variable used to calculate each ordination axis indicate the contribution of that variable to observed object dispersion along that axis. Constrained ordination techniques can be seen as testing specific hypotheses of how environmental variables determine response variable values. Visualization of constrained ordination analysis allows interpretation of these relationships and reveals object similarity. This group of methods also includes several multivariate statistical tests that are used to assess whether the observed data distribution can be expected due to chance.
3. Discriminatory methods are an extension of the interpretive multivariate techniques and are usually called discriminant analyses (DA). The goal of DA is to define discriminant functions (synthetic variables) or hyperspace planes that will maximize the separation of objects among different classes (Box 4). For example, in a linear discriminant analysis (LDA), the measured variables serve as a set of explanatory variables, and the response variable defines the class of each object. The discriminant function(s) are constrained to be specific combinations of explanatory variables. Variable coefficients (often called weights or loadings) used to calculate each discriminant function indicate the relative contribution of each of the explanatory vari-

Box 3. Pitfalls of the use of compositional data

Many types of biological and ecological data are expressed as relative-to-total values, such as relative abundances of different microbial taxa in a soil site, or fraction of particular metabolite in the metabolic profile of a sample. The widespread use of relative values stems from the technical limitations of employed experimental methods, which do not allow direct comparison of the raw measured values of a particular variable (such as gene mRNA level or microbe abundance) between samples. In many cases, an assumption is made that the overall measured signal should be the same among different samples, and the levels of variables are expressed as a fraction of the total. Unfortunately, such transformation of raw data into relative form gives rise to the phenomenon of constant-sum constraint, which violates the assumption of variable independence made in many statistical tests. Specifically, because the sum of all values in a sample has to equal a predefined value (for example, 1% or 100%), a change of one of the variables in that sample will cause reciprocal changes in the calculated *relative* values for other variables (Faust *et al.* 2012). This will produce a negative correlation effect between such variables (see Box Table). If the assumption of the equal overall sum of all measured values among objects is indeed true, such use of compositional data can lead to biologically valid interpretations. However, if 'equal sum' assumption can only be technically but not biologically justified, the use of compositional data can produce false discoveries (see Box Table), as the joint probability distribution of compositional variables cannot describe the distribution of the underlying absolute variables (Lovell *et al.* 2015). Ordination techniques can similarly produce erroneous results because many of these techniques are based on the calculation of the matrix of pairwise distances or dissimilarities which are often a function of correlation (Lovell *et al.* 2015). Rank-based nonparametric statistical measures can diminish the effects of outliers, random noise and deviations from expected probability distribution on the correlation estimates (Shevlyakov & Smirnov 2011). They, however, still suffer from the possible lack of congruency between absolute and relative data distributions as shown by Lovell *et al.* (2015).

Several approaches are available to mitigate the described issues in compositional data analysis.

1 A centred log-ratio transformation can be applied to the compositional data set as $x'_i = \log(x_i/(\prod x_i)^{1/n})$, where $(\prod x_i)^{1/n}$ is the geometric mean of all variables x_i ($i = 1\ldots n$) in an object (Aitchison 1986; Lovell *et al.* 2015). This transformation removes effect of constant-sum constraint on the covariance and correlation matrices; such log-ratio transformed data set can be subjected to multivariate analyses such as PCA (Kucera & Malmgren 1998). The use of this transformation is, however, challenging for data sets with many zeroes (Friedman & Alm 2012), and a zero-replacement procedure has been proposed (Aitchison 1986).

2 The variance of log-ratio transformed data was used as a basis for the SparCC (*Spar*se *C*orrelations for *C*ompositional data) approach of correlation network inference (Friedman & Alm 2012). SparCC was shown to be very accurate on simulated data and was applied to high-throughput sequencing data sets from Human Microbiome Project to reveal robust taxon–taxon interaction networks.

3 A specific method to allow construction of correlation networks for ecological compositional data, called CCREPE (*C*ompositionality *C*orrected by *RE*normalization and *PE*rmutation), has also been recently developed (Gevers *et al.* 2014). Taking into account the constant-sum constraint, CCREPE adjusts *P*-values assigned to similarity measures (such as correlation coefficient) through a *Z*-test comparison of the observed similarity distribution with the null distribution generated through permutation and renormalization of the data. This approach reduces the number of spurious correlations and false discoveries.

Box Table

Absolute values

	$V1_{raw}$	$T2_{raw}$	$T3_{raw}$	$T4_{raw}$	$T5_{raw}$	$T6_{raw}$	average R_P V-vs-T
S1	10	10	11	10	9	10	-0.07
S2	30	11	10	11	9	9	average R_s V-vs-T
S3	50	10	9	9	10	10	-0.16
S4	70	9	9	9	9	9	
S5	110	11	11	9	11	9	

Relative values

	$V1_{comp}$	$T2_{comp}$	$T3_{comp}$	$T4_{comp}$	$T5_{comp}$	$T6_{comp}$	average R_P V-vs-T
S1	0.17	0.17	0.18	0.17	0.15	0.17	-0.99
S2	0.38	0.14	0.13	0.14	0.11	0.11	average R_s V-vs-T
S3	0.51	0.10	0.09	0.09	0.10	0.10	-1.00
S4	0.61	0.08	0.08	0.08	0.08	0.08	
S5	0.68	0.07	0.07	0.06	0.07	0.06	

S1…S5 – individual samples; V – variable with varied abundance; T- stable variable (10±1); R_P - Pearson correlation coefficient; R_s - Spearman correlation coefficient

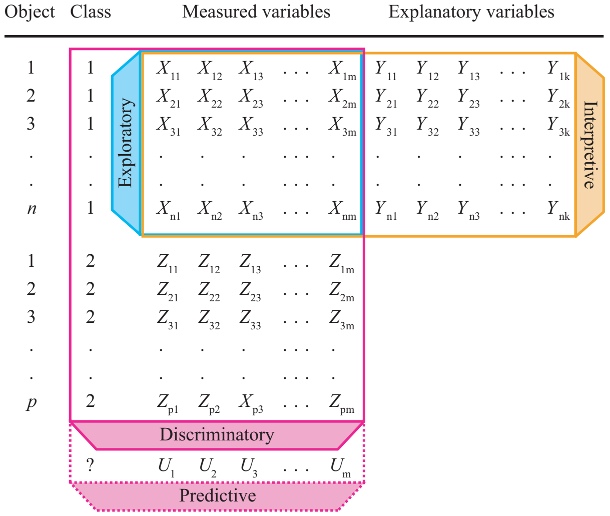
ables to the observed object separation along each discriminant function. This can provide biological insight into the determinants of the (dis)similarity among object classes. Using the calculated discriminant functions, DA algorithms also generate prediction models that permit predictive classification of new objects into one of the classes based on the values of the measured explanatory variables.

This review aimed to provide a short description of the most widely used multivariate statistical approaches in microbial ecology. These methods are listed in Table 1, together with the classification of each method, expected data structure, and availability of MATLAB and R scripts to run each algorithm. Figure 1 provides an illustration of the input expectations for each category of multivariate techniques. A more in-depth description of these and other multivariate analyses can be obtained from available literature (Manly 2004; Borcard *et al.* 2011; Legendre & Legendre 2012; James *et al.* 2014; Šmilauer & Lepš 2014) and online sources such as GUSTA ME (Buttigieg & Ramette 2015). Stand-alone software packages such as CANOCO (Šmilauer & Lepš 2014), PRIMER v6 (Clarke & Gorley 2005) and PAST (Hammer *et al.* 2001) are available as alternatives to MATLAB and R scripts.

Exploratory methods

Principal components analysis

Principal components analysis (PCA) is one of the most widely used and one of the oldest methods of ordination analyses (Pearson 1901). The general principle of PCA is to calculate new synthetic variables called principal components based on the matrix operations applied to the original data set of quantitative variables. Each principal component (PC) is a linear combination of original variables calculated so that the first PC represents an axis in the multidimensional data space that would produce the largest dispersion of values along this component (see Box 4 Fig. A). Other principal components are calculated as orthogonal to the preceding components and similarly are positioned along the largest remaining scatter of the values. Thus, PCA creates a rotation of the original system of coordinates so that the new axes (principal components) are orthogonal to each other and correspond to the directions of largest variance in the data set.



Object	Class	Measured variables					Explanatory variables				
1	1	X_{11}	X_{12}	X_{13}	. . .	X_{1m}	Y_{11}	Y_{12}	Y_{13}	. . .	Y_{1k}
2	1	X_{21}	X_{22}	X_{23}	. . .	X_{2m}	Y_{21}	Y_{22}	Y_{23}	. . .	Y_{2k}
3	1	X_{31}	X_{32}	X_{33}	. . .	X_{3m}	Y_{31}	Y_{32}	Y_{33}	. . .	Y_{3k}
.	.	.	.	.	. . .	.	.	.	.	. . .	.
.	.	.	.	.	. . .	.	.	.	.	. . .	.
n	1	X_{n1}	X_{n2}	X_{n3}	. . .	X_{nm}	Y_{n1}	Y_{n2}	Y_{n3}	. . .	Y_{nk}
1	2	Z_{11}	Z_{12}	Z_{13}	. . .	Z_{1m}					
2	2	Z_{21}	Z_{22}	Z_{23}	. . .	Z_{2m}					
3	2	Z_{31}	Z_{32}	Z_{33}	. . .	Z_{3m}					
.	.	.	.	.	. . .	.					
.	.	.	.	.	. . .	.					
p	2	Z_{p1}	Z_{p2}	X_{p3}	. . .	Z_{pm}					
	?	U_1	U_2	U_3	. . .	U_m					

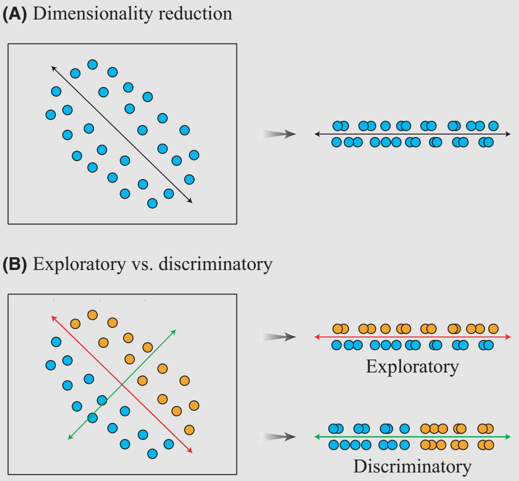
Fig. 1 Expected input data structure for the multivariate ordination techniques. Exploratory techniques (blue rectangle) are used to discern patterns within a single $m \times n$ matrix of measured variables and objects. Interpretive techniques (orange rectangle) are used to explain variation in a set of dependent variables (measured variables) by another set of independent variables (explanatory variables). Discriminatory techniques (pink rectangle) are used to separate objects between different classes based on the values of measured variables. Predictive approaches are an extension of discriminatory techniques that allow classification of a new object based on the generated discrimination model.

By definition, the first PC axis of the PCA output represents the largest gradient of variability in the data set, PC2 axis—the second largest, and so forth, until all data set variability has been accounted for. Each object can thus be given a new set of coordinates in the principal components space, and the distribution of objects in that space will correspond to the similarity of the variables' scores in those objects (Box 4). Displaying objects in only the first two or three PC dimensions is often sufficient to represent much of the variability in the original data set. In fact, looking at the per cent of data set variance that is 'explained' by each principal component can tell us whether there are any dominant gradients in the data set. We can also infer whether few (first few PCs explain almost all variance) or many (variance 'explained' by each successive PC is distributed along a shallow curve) effects influenced the observed data distribution.

Because PCA uses Euclidean distance to measure dissimilarity among objects, care should be taken when using PCA on a data set with many zeroes, as is often the case for data with long gradients. As described in detail by Legendre & Gallagher (2001), when run on such data sets, PCA can generate severe artefacts such as horseshoe visualization effect (see Legendre & Legendre 2012; ter Braak & Šmilauer 2015, for examples). With this artefact, objects at the edges of the environmental gradient actually appear close to each other in the ordination space (Novembre & Stephens 2008). While the horseshoe effect can be partially reduced by processing of the original data values through a chord or Hellinger transformation before running PCA (Box 2; Legendre & Gallagher 2001), the use of correspondence analysis (CA) is usually advocated for such data sets (ter Braak & Šmilauer 2015).

Principal components analysis can be used as a simple visualization tool to summarize data set variance and show the dominant gradients in low-dimensional space. PCA results are usually displayed as a two- or three-dimensional scatter plot, where each axis corresponds to a chosen principal component, and each object is plotted based on its corresponding PC values.

Box 4. Introduction into ordination techniques


(A) Dimensionality reduction

(B) Exploratory vs. discriminatory

Ordination methods are a group of multivariate statistical approaches that attempt to order objects based on the values of variables measured in these objects (Goodall 1954). Ordination methods are also often called gradient analyses, because these algorithms look for the main gradients of variation in multidimensional data space, and then arrange the objects in a new system of coordinates with each gradient serving as an axis as shown in Fig. A. Often, the few largest gradients account for the majority of the variance in the data. Thus, a large data matrix can be displayed in just 2 or 3 dimensions with relative positioning of the objects representative of the relationships among the variables measured in these objects. This dimensionality reduction coupled with the straightforward interpretation of visual distribution of objects on an ordination plot (close objects have similar values of variables, distant objects—different) make these ordination techniques a popular choice in ecological studies.

Exploratory ordination methods look for the largest gradients of the scatter in the data, and object distribution in the ordination space represents the largest sources of dispersion (red arrow in Fig. B). Examination of this output can reveal groups of similar objects, and it can show whether the main gradients of variance correspond to the expected sources of data variation. Alternatively, if the potential sources of data variation have been measured, these additional predictor variables can then be used to guide the analyses using one of the 'constrained' ordination techniques. In these methods, the goal is to look specifically for gradients in the multidimensional data space that are associated with the observed variation of predictor variables. If the additional variable describes the distribution of objects among different classes, discriminatory ordination analysis will maximize the separation of objects belonging to different classes (green arrow in Fig. B).

Often, there are many different predictor variables that all influence measured response variables. This can make interpretation of ordination results difficult as it is harder to estimate the contribution of each predictor variable to the observed object distribution (see Fig. 6A for an example). To more specifically assess the relationship between particular predictor variable(s) and response data set, partial ordination analysis can be performed (ter Braak 1988). In this approach, one or few predictor variables are chosen, and the rest are treated as covariates. Contribution of covariates to the variability in response variables is factored out, and ordination is then performed on the residual variability. Partial ordination can reveal how much of the total variation in response variables is attributed specifically to the chosen predictor variable(s), and it can remove the effects of undesired or accidental predictors (Liu 1997). Partial ordination can be extended further and used as a basis of variation partitioning (Borcard *et al.* 1992). This approach allows one to partition the total variation in the data set of response variables into components uniquely as well as jointly described by individual (groups of) predictor variables. The total variation is then displayed as pie chart or Venn diagram, with each section representing the fraction of total variance explained by each (group of) predictor(s). This reveals which predictors have the highest effect on the response variables, and it also shows any joint effects between predictor variables. Volis *et al.* (2011) and Hildebrand *et al.* (2013) provide examples of variation partitioning analysis.

While it is typical to show PC1-vs.-PC2 and/or PC2-vs.-PC3 scatter plots, any two or three principal components can be chosen for visualization. Examples of PCA use in microbial ecological studies are provided by Hong *et al.* (2010), Ringel-Kulka *et al.* (2013) and Shankar *et al.* (2013).

Correspondence analysis

Correspondence analysis is an exploratory technique designed to find relationships (correspondence) between rows and columns of a matrix of tabulated data (often called contingency table) and to represent these relationships in an ordination space (Hill 1973, 1974). While the goal of CA is similar to that of PCA, important differences exist between these methods:

1 PCA maximizes the amount of explained variance among measured variables, while CA maximizes the correspondence (measure of similarity of frequencies) between rows (represent measured variables) and columns (represent objects) of a table (Yelland 2010).

Thus, transposing the data matrix prior to running CA produces the same result, while this is not true for PCA (Choulakian 2001). This property leads to CA results ordinarily displayed in a biplot, where both row (variables such as taxa, genes, metabolites) and column (objects such as samples, sites) variables are jointly depicted on the same ordination chart. For each variable, the position of the point on a plot represents its estimated optimum along the gradient.

2 While PCA expects a linear relationship among variables, CA is based on a unimodal model. This is a very important distinction as we describe in detail in Box 5, and the choice between PCA and CA should depend on the type of variables being analysed, assumed response model, and the lengths of the gradients and variable distributions (Šmilauer & Lepš 2014).

3 The distances among objects in full CA ordination space are equal to a variant of Euclidean distance, called the weighted Euclidean distance or chi-square distance. The calculation of chi-square distance does not take into account cases where the value of a variable in two different objects is zero. Thus, CA is well suited to the analysis of data sets of species composition and abundance, where sites at both ends of environmental gradient usually lack many common species. Note, however, that as described in detail by ter Braak & Šmilauer (2015), true chi-square distance between objects is only visualized in the full CA ordination space. In two dimensions and with detrending applied, the interobject distances now represent actual ecological distances (ter Braak & Šmilauer 2015).

While CA does not suffer from the horseshoe effect described above for PCA, unmodified CA ordination output can often produce a noticeable 'arch' effect (see ter Braak & Šmilauer 2015, for an example). Arch effect is a mathematical artefact where distribution of objects along the first canonical axis is also partially mirrored in the second axis, because CA axes are constrained to be uncorrelated but not necessarily unrelated of each other (Hill & Gauch 1980). The arch effect can be eliminated by the use of a detrended correspondence analysis (DCA) technique, which attempts to remove residual axis 1 variance from axis 2 object scores (Hill & Gauch 1980; Lockyear 2000; ter Braak & Šmilauer 2015). Detrending can also correct the shrinking of interobject distances at the ends of ordination axes and can reduce the influence of uncommon variables (e.g. rare species). Note that while often useful, the detrending procedure does not always lead to an ordination output that better reflects the observed environmental gradients (see Legendre & Legendre 2012, for an in-depth discussion).

Correspondence analysis and DCA were shown to approximate well the ecological niche model and the typically expected Gaussian distribution of ecological variables (ter Braak & Looman 1986; ter Braak 1987). These techniques have been applied in ecology to analyse species distribution across many sites and samples, and to visualize sample similarity based on the presence and absence of different species (Prideaux *et al.* 2013). Other examples of CA use can be found in Jones *et al.* (2007), Perez-Cobas *et al.* (2014) and Thureborn *et al.* (2013).

Principal coordinates analysis

Principal coordinates analysis (PCoA) is a conceptual extension of the PCA technique described above. It similarly seeks to order the objects along the axes of principal coordinates while attempting to explain the variance in the original data set. However, while PCA organizes objects by an eigen analysis of a correlation or covariance matrix, PCoA can be applied to any distance (dissimilarity) matrix (Gower 1966). PCoA has gained recent popularity in microbial ecology due to its ability to use phylogenetic distance (e.g. UniFrac distance; Lozupone & Knight 2005) and community composition (e.g. Bray–Curtis distance; Bray & Curtis 1957) measures to calculate (dis)similarity among microbial populations. Another potentially good choice of a distance measure is a recently introduced distance correlation metric *dCor*, which is able to robustly capture nonlinear relationships between variables (Szekely *et al.* 2007). Because PCoA uses distance matrix as its input, it is not possible to directly relate any of the measured variables to individual principal coordinate axes (Ramette 2007). An indirect correlation or regression analysis of object PC values vs. object scores for a particular variable can instead be used to estimate the contribution of that variable to object dispersion along a particular PC axis (Koenig *et al.* 2011).

A particularly useful implementation of PCoA for the microbial ecology studies has been developed by Rob Knight group (Lozupone & Knight 2005). This PCoA uses a measure of phylogenetic community distance, called UniFrac metric, calculated as the fraction of total branch lengths on a community phylogenetic tree that are unique to one microbial population or the other. Thus, if the two populations are identical and the same members are found in both, all the phylogenetic tree branches are shared, and the UniFrac value is 0. If no member is shared by two populations, then all of the branches are unique, thus providing the maximum possible UniFrac distance. The calculated UniFrac distance metric can therefore be used as a measure of the phylogenetic similarity of the community structure between

Table 1 Common multivariate statistical tools

Technique	Assumed relationship[*]	Input	R script[†]	MATLAB script[†]
Exploratory				
PCA	Linear	Raw data	prcomp (stats)	princomp (built-in)
CA/DCA	Unimodal	Raw data	ca (mva) decorana (vegan)	CAR[‡]
PCoA	Any[DM]	Distance matrix	pcoa (ape)	f_pcoa (Fathom toolbox[§])
NMDS	Any[DM]	Distance matrix	metaMDS (vegan)	mdscale (built-in)
Hierarchical clustering	Any[DM]	Distance matrix	hclust (stats)	pdist + linkage + cluster (built-in)
K-means clustering	Any[DM]	Distance matrix	kmeans (stats)	kmeans (built-in)
Interpretive				
CCorA	Linear	Raw data	CCorA (vegan)	f_CCorA (Fathom toolbox[§])
CIA	Any[ORD]	Ordination output	coinertia (ade4)	coinertia.m[¶]
PA	Any	Any	procrustes (vegan)	f_procrustes (Fathom toolbox[§])
RDA	Linear	Raw data	rda (vegan)	f_rda (Fathom toolbox[§])
db-RDA	Any[DM]	Distance matrix	capscale (vegan)	f_rdaDB (Fathom toolbox[§])
CCA	Unimodal	Raw data	cca (vegan)	CAR[‡]
PRC	Linear	Raw data	prc (vegan)	—
GLM	Any[LF]	Raw data	glm (stats)	glmfit (built-in)
Mantel test	Any	Distance matrix	mantel (vegan)	f_mantel (Fathom toolbox[§])
ANOSIM	Any	Distance matrix	anosim (vegan)	f_anosim (Fathom toolbox[§])
PERMANOVA	Any	Distance matrix	adonis (vegan)	f_npManova (Fathom toolbox[§])
Discriminatory				
DFA	Linear	Raw data	candisc (candisc)	DA (Classification toolbox[**])
OPLS-DA	Linear	Raw data	oplsda (muma)	osccalc + pls (PLS_toolbox®)
SVM	Any[KF]	Raw data	ksvm (kernlab)	svmtrain (built-in)
RF	Any	Raw data	randomForest (randomForest)	classRF_train (RF_MexStandalone[††])

DM, assumed relationship depends on the distance metric used. ORD, assumed relationship depends on the ordination technique used. LF, assumed relationship depends on the link function used. KF, model can be linear or nonlinear if a nonlinear kernel function is incorporated.

[*] Expected relationship among variables.

[†] Entries indicate functions for the respective platforms. Names within brackets represent the package that contains the function.

[‡] Described in Lorenzo-Seva *et al.* (2009).

[§] Described at http://www.marine.usf.edu/user/djones/matlab/matlab.html and available from: http://www.marine.usf.edu/user/djones/.

[¶] Described in Doledec & Chessel (1994).

[**] Described in Ballabio & Consonni (2013).

[††] Available at https://code.google.com/p/randomforest-matlab/.

populations. A variant of the algorithm, called weighted UniFrac (Lozupone *et al.* 2007), was also introduced, which weighs the tree branches according to the abundance of each community member. Examples of PCoA use in ecological research are provided by Fierer *et al.* (2012), Koren *et al.* (2013) and Schnorr *et al.* (2014).

Nonmetric multidimensional scaling

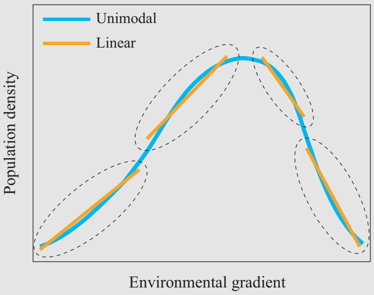
Multidimensional scaling (MDS) is a unique ordination technique in that a (small) number of ordination axes are explicitly chosen prior to the analysis and the data are then fitted to those dimensions. Thus, if only 2 or 3 axes are chosen, there will be no nondisplayed axes of variation at the end of the analysis. Similar to PCoA, a matrix of object dissimilarities is first calculated using a chosen distance metric. In nonmetric multidimensional scaling (NMDS), ranks of these distances among all objects are calculated. The algorithm then finds a configuration of objects in the chosen N-dimensional ordination space that best matches differences in ranks (Kruskal 1964).

Because NMDS is a numerical rather than an analytical technique, it does not produce a unique solution. A 'stress' parameter is computed to measure the lack of fit between object distances in the NMDS ordination space and the calculated dissimilarities among objects. The NMDS algorithm then iteratively repositions the objects in the ordination space to minimize the stress function

Box 5. Models of variable response to environmental gradients

The goal of ecological studies is to assess and contrast the relationships between (groups of) biological entities (species, metabolites, proteins, etc.) and their environment. To accomplish this, parametric statistical techniques need to make a specific assumption of the type of the relationship, generally called variable response model, in order to define the appropriate mathematical calculations. While linear relationships are the simplest (the relationships are modelled by a straight line), true linear gradients are rare in nature (Whittaker 1967). Instead, current evidence shows that biological variables tend to display a nonmonotonic relationship with the environment and have an optimum and low-density areas (Whittaker 1956, 1967). For example, consistent with the ecological niche model, species distribution along a gradient (spatial, temporal, response to perturbation, etc.) usually displays a unimodal shape with higher population density in the middle of the distribution area and fewer members towards the edges of the territory (Whittaker *et al.* 1973; Austin 2007). The standard statistical methods that assume linear relationships are thus not appropriate for the analyses of such variables. Instead, Gaussian models, first promoted by Gauch & Whittaker (1972), were found to be a good approximation of the distribution of ecological variables. These models are also well suited to deal with compositional data, data with many zeroes and binary (presence/absence) variables (ter Braak & Verdonschot 1995). There are, however, two caveats. First, linear-based methods can be used on some nonlinear response data sets if the data can first be transformed to linearity (Box 2). Second, note that over a narrow range of environmental gradient, distribution of response variable can approximate linear function, and thus, linear methods such as PCA might be appropriate (Box Fig. 2). Further discussion of this topic can be found in Austin (2007). An excellent guide for the gradient length-based choice between unimodal and linear-based ordination approaches can be found in Šmilauer & Lepš (2014).

Based on considerations described above, several different types of response models can be chosen. Linear response models are implied in the Euclidean distance-based methods such as PCA and RDA. Unimodal responses can be described explicitly with maximum-likelihood estimators of Gaussian curves, or heuristically with weighted averaging estimators (ter Braak & Looman 1986; ter Braak & Prentice 1988). The ordination methods based on the former approach are available but have generally not been used often due to complexity (Yee 2004). Weighted averaging estimators are much simpler and were used as a basis for the 'reciprocal averaging' method (correspondence analysis) developed by (Hill 1973) as an approximate solution to Gaussian ordination. Finally, a number of ordination methods such as PCoA, db-RDA and NMDS are not based on specific underlying model of variable–environment relationship, but instead rely on the user-defined matrix of (dis)similarities among all objects.



(Dugard *et al.* 2014). Stress values ≤0.15 are considered generally acceptable (Clarke 1993).

Nonmetric multidimensional scaling makes few assumptions about the distribution of data, and is often used for molecular fingerprinting techniques such as denaturing gradient gel electrophoresis (DGGE) and terminal restriction fragment length polymorphism (TRFLP) (Littman *et al.* 2009). For data sets with many different gradients of variance, NMDS ordination can be superior to that of other ordination techniques (Minchin 1987). That is because with $N_{\text{axis}} = 2$ all data set variance is utilized to distribute objects in a two-dimensional NMDS ordination plot, whereas the first two dimensions of the PCA/CA/PCoA ordination only display a part of that variance (Legendre & Legendre 2012). NMDS, however, does not perform a simultaneous ordination of both variables and objects. Note that NMDS is not an eigenvector-based gradient analysis technique but rather is a mapping method. Each of its ordination axes does not correspond to a particular gradient in the original data set, and its goal is to represent ranks of pairwise dissimilarities among objects. Examples of how NMDS has been utilized in ecological research can be found in Mason *et al.* (2014) and Ushio *et al.* (2015).

Cluster analysis

The goal of CLA is to separate variables into groups based on the similarity of variables' scores among objects, so that variables within each group (cluster) are more similar to one another than to variables in other groups (Driver & Kroeber 1932). The algorithms minimize the within-group distances and maximize between-group distances. The same approach can also be used to distribute objects into groups. After objects/variables are clustered into individual groups, all members of the same group can be considered together. This facilitates the interpretation of high-throughput data. CLA has been particularly popular in the analyses of high-throughput microarray gene expression data, because it is more straightforward to examine just a few different types of gene expression responses after environmental perturbation instead of analysing each individual gene behaviour separately (Withman *et al.* 2013). Because CLA allows the use of any distance metric that can generate (dis)similarity measures, this approach can be equally well suited to ecological data (Gajer *et al.* 2012; Shankar *et al.* 2014). Below, we describe the two most common types of CLA.

Hierarchical clustering. Hierarchical clustering (HCA) produces a joint treelike organization of variables, with organization and length of branches indicative of the relative similarity of different variables. HCA is usually accomplished through an agglomerative approach where single variables that are most similar to each other are first successfully joined into common nodes, which are then joined with other nodes and so forth. This process continues iteratively until all variables have been incorporated into nodes that form a single connected structure. The order of how the nodes were formed and joined is displayed through the organization of the nodes into a 'hierarchy' tree (called dendrogram). The connectivity and length of tree branches reflect the similarity between variables and nodes. There are several different methods to compute the distances between nodes (e.g. single linkage, complete linkage and average linkage, the latter is generally used most often), which influences relative node positioning and tree branching (Ferreira & Hitchcock 2009). HCA provides a simple way to visualize similarities among variables, and the dendrogram structure can be used to make inferences about grouping of variables. A special case of HCA is biclustering (sometimes called two-way clustering), which produces simultaneous clustering of rows and columns of the data matrix. Biclustering can find features (genes, microbial taxa, etc.) that correlate only in a subset of objects but not in the rest of the data set (Sridharan *et al.* 2014; Shankar *et al.* 2015).

Disjoint clustering. We use this term to describe a number of related clustering techniques that aim to separate all variables or objects into individual, usually mutually exclusive, and in most cases unconnected clusters. One of the most frequently used disjoint clustering methods, K-means clustering, seeks to partition all variables into predefined K number of individual clusters. The initial cluster reference profiles (variable's value for each object) are either generated randomly, defined by user, or a single random variable is chosen in the beginning to seed each cluster. All variables are then assigned to one of these predefined clusters based on the shortest distance of each variable to cluster centroids. Each cluster reference profile is then recalculated based on the mean of the variables in that cluster, and all variables are repartitioned into clusters. This process is repeated until a stable solution is achieved. Because initial cluster reference profiles are often chosen randomly, K-means clustering is considered nondeterministic, so repeats of this procedure on the same data set can produce somewhat different results. Thus, it is advisable to run K-means algorithm multiple times and then choose the clustering result that achieves the lowest total error sum of squares (sum of squared distances among variables within each cluster; represents how 'tight' variables are within each cluster; Arthur & Vassilvitskii 2007). The chosen number of K clusters has a large impact on cluster profiles, and there are mathematical approaches to define the optimal number of clusters (Hastie *et al.* 2001). K-means clustering was used to define groups of genera that are modified by the faecal microbiota transplantation (FMT) in patients with *Clostridium difficile*-associated disease (Shankar *et al.* 2014), and to associate glycan degradation patterns with microbial abundances in mammalian gut (Eilam *et al.* 2014). Other clustering techniques are available such as TWINSPAN (Hill 1979), self-organizing maps (Kohonen 1982) and DBSCAN (Ester *et al.* 1996).

Interpretive methods

Interpretive methods described in this section can be further subdivided into three types. Symmetric approaches compare two data sets and do not distinguish variables between explanatory and response. These include canonical correlation analysis (CCorA), co-inertia analysis (CIA) and procrustes analysis (PA). Asymmetric approaches also use two different sets of variables but designate one set as explanatory (independent) variables and another as response (dependent) variables. Discussed asymmetric techniques are redundancy analysis (RDA), canonical CA, principal response curves (PRC) and generalized linear models. Finally, we also include in this group three techniques that are focused on the statistical significance testing of multivariate data sets.

Canonical correlation analysis

Canonical correlation analysis is a multivariate extension of simple correlation analysis. The goal of this method is to investigate the associative relationship between two sets of variables. If we have two sets of variables, $X = (x_1, x_2, \ldots, x_n)$, and $Y = (y_1, y_2, \ldots, y_n)$, and there are correlations among the variables, then CCorA will aim to find linear combinations of the x's and the y's that would provide maximum correlation between X and Y. As the goal is to find correlations, there is no assumption of which variables are predictive and which are responsive. CCorA output is a set of orthogonal canonical variates with a corresponding set of canonical correlations. The first canonical correlation is between the first canonical variates X_{CV1} and Y_{CV1} and has the largest value, the second—between the second canonical variates and has the second largest value, and so forth. Once the canonical variates are calculated, we can assess how each original variable contributed towards each canonical variate based on the weight coefficient of that variable. This is usually done for X_{CV1} and Y_{CV1} variates, as they show the strongest correlation in comparison with other pairs. The level of overall association between two variable sets can be assessed by the fraction of their joint covariance explained by all pairs of canonical variates. Examples of the use of CCorA in microbial ecology can be found in Schwartz *et al.* (2012), Wang *et al.* (2012b) and Guan *et al.* (2013).

Co-inertia analysis

The goal of CIA, similarly to CCorA, is to find the strongest associations between two sets of variables. The method was developed by Doledec & Chessel (1994) to specifically study species–environment interactions. In contrast to CCorA, the relationships are based on the covariance rather than on a correlation. In a typical CIA approach, both sets of variables are first subjected to a gradient analysis such as PCA or CA. Using each ordination output, an axis of variance (inertia) is then found in each ordination space so as to achieve the maximum covariance between the projected object values along each axis. Further pairs of axes that maximize the remaining covariance can be calculated under orthogonality constraints. Co-inertia is thus a measure of the similarity of object distribution in both ordination spaces and is quantitatively defined by an *RV* coefficient (Heo & Gabriel 1998; Dray *et al.* 2003a). CIA allows each variable set to be analysed by a different ordination algorithm (Dray *et al.* 2003a). CIA has been used in several recent studies to reveal relationships between human-associated microbiome and metabolome data sets in the gut of the elderly (Claesson *et al.* 2012), on the intestinal mucosal surface (McHardy *et al.* 2013), and in obese subjects (Zhang *et al.* 2015).

Procrustes analysis

Procrustes analysis is a statistical method of comparing the distributions of multiple sets of corresponding objects (Hurley & Cattell 1962; Gower 1975). Because each data set can be depicted as a cloud of objects in a multidimensional space, this analysis is also referred to as comparison of shapes. The aim of this technique is to superimpose structures and then move, rotate, and scale them so as to achieve the best match (the smallest difference in shapes). In microbial ecology studies, PA is usually used to compare the distributions of the same set of objects in different ordination spaces. Here, PA minimizes the square root of sum of squared distances (sometimes called Procrustes distance) between the positions of the same object in different ordination outputs. For example, we have carried out microbial and metabolite profiling of the same set of faecal samples obtained from several cohorts of children (Shankar *et al.* 2013, 2015). Several different ordination analyses indicated that samples could be separated largely based on their cohort assignment in both microbiota and metabolite data sets. By employing PA of the distributions of samples in PCA plots based on these two data sets, we showed that the separation of samples was congruent between metabolite and microbial data sets. Position of a particular sample in microbiota-based ordination plot was generally close to its placement in the metabolite-based ordination chart (Fig. 2). By randomly changing the object IDs in one of the sets, a statistical significance of the observed minimized Procrustes distance was obtained (Shankar *et al.* 2015). Similar to CIA, PA can be applied to outputs from any ordination method. Thus, it can also be used to assess whether multiple ordination techniques applied to the same object-by-variable data set produce similar results (Bassett *et al.* 2015). Other examples of PA use are provided in studies by Claesson *et al.* (2012), McHardy *et al.* (2013) and Zhang *et al.* (2015).

All three symmetrical techniques described above are closely related. Compared to CCorA, both CIA and PA impose no constraints on the number of variables in the data sets compared to the number of objects, and they have fewer assumptions (Legendre & Legendre 2012). CIA provides a more readily interpretable quantitative assessment of the strength of the global association (*RV* coefficient), while PA can be run on a wider range of input tables. Both CIA and PA can be applied to more than 2 data sets simultaneously (Ten Berge 1977; Bady *et al.* 2004). An approach merging CIA and PA within the same analysis has been proposed (Dray *et al.* 2003b).

Redundancy analysis

Redundancy analysis is a type of constrained ordination that assesses how much of the variation in one *set* of variables can be explained by the variation in another *set* of variables. It is the multivariate extension of simple linear regression that is applied to *sets* of variables (Rao

1964; Van den Wollenberg 1977). RDA is based on similar principles as PCA, and thus assumes linear relationships among variables. RDA is in fact a canonical version of PCA where the principal components are constrained to be linear combinations of the *explanatory* variables. If the expected relationship between response variables and environmental gradients is unimodal rather than linear (Box 5), then the canonical CA described below is more appropriate.

A combination of two data sets is required to run RDA: the first data set contains response (dependent) variables (species presence or abundance, metabolite levels, etc.) and the other set contains explanatory (predictive) variables (such as environmental variables or experimental treatments measured in the same samples or sites) (see Fig. 1). 'Redundancy' expresses how much of the variance in the set of response variables is explained by the set of explanatory variables. The fraction of the total variance observed in response variables that is explained by all the explanatory variables is a useful indication of how much variance in the species distribution, for example, is due to differences in environmental factors between sites. The output of RDA is an ordination that is usually shown on a two-dimensional 'triplot', with constrained RDA dimensions used as axes. Each object is depicted by a point, and response variables are represented by arrows originating from the coordinate system origin, and explanatory variables by either arrows (quantitative variables) or points (categorical variables). Because triplots display a lot of data on a single plot, their interpretation is more challenging. On a distance triplot, distances between objects represent the between-object similarity; the angles between arrows of response variables and arrows of explanatory variables represent the found associations between those variables. Finally, the projection of an object onto an arrow approximates the value of the corresponding variable in this object. A detailed description of the interpretation of ordination diagrams can be found in ter Braak & Verdonschot (1995) and Šmilauer & Lepš (2014). Alternatively, to simplify interpretation, one can carry out partial constrained ordination such as PRC analysis described below (see also Box 4).

A special variant of RDA, called distance-based RDA or db-RDA, can be used when the response data are available as dissimilarity matrix, or when the use of Euclidean distance is inappropriate (Legendre & Anderson 1999; Anderson & Willis 2003). db-RDA is thus a constrained version of PCoA, and it provides an opportunity to use phylogenetic and ecological distances in constrained ordination analysis (Shakya *et al.* 2013). Examples of the use of RDA for constrained analysis of ecological data sets are available in the works by Rajilic-Stojanovic *et al.* (2010), Zhang *et al.* (2012a) and Ringel-Kulka *et al.* (2013).

Canonical correspondence analysis

The goal of the canonical correspondence analysis (CCA) is similar to that of RDA as it too aims to find the relationship between two *sets* of variables X and Y. However, whereas RDA assumes a linear relationship among variables, CCA expects a unimodal relationship. Thus, CCA is a canonical form of CA of the response variable set that is constrained by the set of explanatory variables (ter Braak 1986). Note that CCorA can also be

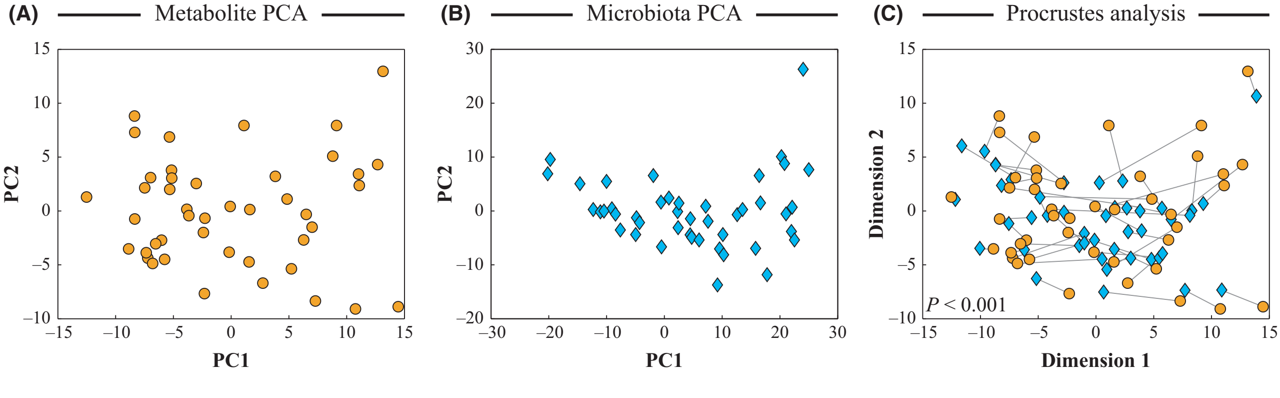


Fig. 2 Use of procrustes analysis to test for congruency between ordination outputs. Principal components analysis (PCA) was performed on centred log-ratio transformed compositional metabolite binned data (panel A) and compositional microbiota phylotype data (panel B) for the same set of human faecal samples. Raw metabolite and microbiota data sets were taken from Shankar *et al.* (2015). PCA results were provided as input to procrustes analysis to compare object positioning in PCA ordination spaces. Panel C displays the first two dimensions of the procrustes analysis output. The distance between each sample position on two PCA plots is indicated by connecting line. Shorter lines represent more similar object positioning on both PCA plots. *P*-value for statistical significance of observed object separation congruency was generated using randomization of object labels.

abbreviated as CCA in the literature. These are, however, two different techniques. The visual output of CCA is a triplot similar to that of RDA, where in addition to objects and variables typically shown on a CA plot, the explanatory variables are also shown as either arrows (quantitative variables) or points (categorical variables). An example of CCA plot interpretation is provided below. Examples of CCA utilization in ecological studies can be found in Wang *et al.* (2012a), Zhang *et al.* (2013) and Yan *et al.* (2015).

Principal response curves

When data sets contain a large number of explanatory and response variables, interpretation of RDA and CCA plots can become quite challenging, especially when there are interaction effects among explanatory variables (ter Braak & Šmilauer 2015). In such cases, an influence of any individual explanatory variable on the observed distribution of objects is not readily identified. To facilitate visual representation of such complex constrained analyses and to limit interpretation to a single effect (single explanatory variable) and an interaction with a second effect, principle response curves method can be used [not to be confused with a distinct 'principal curves' technique proposed by De'ath (1999)]. PRC was initially developed by van den Brink & ter Braak (1999) to visualize and interpret differences between treatment and control communities over time. PRC algorithm carried out a partial RDA ordination (Box 4) to partition the variance between individual effects and their interaction term. The canonical coefficients obtained in RDA for each sample were then displayed in a line chart (principal curve) with time plotted on the *X* axis (van den Brink & ter Braak 1999). PRC is especially useful for the analysis of longitudinal series of measurements, as time-dependent effects can be clearly isolated in PRC from other effects (Fig. 6B). PRC was further extended by van den Brink *et al.* (2009) to use a single reference point as control for each series of measurements, allowing longitudinal analysis of individual communities. Multiple sets of objects can be displayed on the same chart as separate response curves, and time can be substituted by other gradient present in the experimental design or the data set (van den Brink *et al.* 2009; ter Braak & Šmilauer 2015). The congruency between each variable response pattern and computed principal response curve is provided by variable weights, which are displayed on a separate chart (Fig. 6B; van den Brink & ter Braak 1999). Although PRC is yet to be extensively utilized in microbial ecology research, few examples are provided in the reports by Zhang *et al.* (2012b) and Fuentes *et al.* (2014).

Generalized linear modelling

Generalized linear modelling (GLM) is a term that describes a statistical approach to relate response variable(s) to the linear combinations of the explanatory (predictor) variable(s) (Nelder & Wedderburn 1972). GLM is an extension of the standard linear models such as regression and analysis of variance (ANOVA). The power of GLM lies in its ability to generate not only regression models for continuous response variables, but also models for discrete and categorical response variables (Nelder & Wedderburn 1972). In GLM, the values of response variable are 'predicted' from a linear combination of explanatory variables by connecting them via a so-called *link* function. Many different link functions can be chosen (examples include log, inverse, power, root and logit functions), and several different distributions of the response variable may thus be defined (examples include linear, Gaussian, logistic and beta distributions). This provides GLM with a tremendous flexibility. The output of GLM is (i) a set of regression coefficients defining the modelled relationship between explanatory and response variables, and (ii) statistical assessment of the fit of the model to the data (Fox 2008). The generated model can then be evaluated to reveal the strength and statistical significance of the relationship between each predictor variable and the response variable. Multivariate extension of GLM is available (Warton 2011), an ordination algorithm based on GLM dimensionality reduction has been developed (Yee 2004), and further extension of GLM, called generalized additive modelling, has been made (Hastie & Tibshirani 1986). Several recent studies have indicated that for ecological data sets that often contain many zeroes, GLM carried out on the data set of measured variables performed better than standard parametric analyses conducted on log- or power-transformed values (O'Hara & Kotze 2010; Warton *et al.* 2012).

Mantel test, analysis of similarities (ANOSIM) and permutational multivariate analysis of variance (PERMANOVA)

Mantel test, ANOSIM and PERMANOVA are multivariate statistical tests of significance. Mantel test typically compares two distance matrices that were calculated for the same set of objects but that are based on two independent sets of variables (e.g. a species dissimilarity matrix and site distance matrix) (Mantel 1967). The test calculates the correlation between values in the corresponding positions of two matrices. Significance of the linear relationship between matrices is assessed through permutation of objects (Box 6). The goal of Mantel test is similar to that of CCorA, CIA and PA (Lisboa *et al.* 2014).

ANOSIM tests for significant difference between two or more classes of objects based on any (dis)similarity measure (Clarke 1993). It compares the ranks of distances between objects of different classes with ranks of object distances within classes. The basis of this approach is similar to the NMDS ordination technique described above. As ANOSIM is based on ranks, it has fewer assumptions compared to the regression techniques such as multivariate analysis of variance (MANOVA).

PERMANOVA is a nonparametric method to conduct multivariate ANOVA and test for differences between object classes (Anderson 2001). Any dissimilarity metric can be used, and the test statistic is calculated from the comparison of dissimilarities among interclass objects to those among intraclass objects. Significance levels (*P*-values) are obtained through permutation (Box 6).

Note that most ordination techniques can also assess the statistical significance of observed object distribution via permutation-based analysis (Box 6).

Discriminatory methods

Discriminant function analysis

Discriminant function analysis (DFA) comprises a group of ordination techniques that find linear combinations of observed variables that maximize the grouping of objects into separate classes (Fisher 1936). Here, the measured variables are the predictor variables, and the variable that defines object classes is treated as the response variable (also called grouping variable). In literature, DFA is also called either LDA, or canonical discriminant analysis (CDA, also called multiple discriminant analysis; usually implies that more than two classes of objects are available). Algorithmically, DFA methods generate latent variables (called discriminant functions or DFs) that maximize the formation of coherent, well-separated clusters of objects. Multiple discriminant functions can be extracted, each orthogonal to the others, until their number equals the number of predictor variables or the number of object classes minus one, whichever is smaller. An eigenvalue associated with each DF defines the discriminating 'power' of that function. DFA techniques are related to PCA; however, unlike PCA which summarizes total variation in the data set, LDA and CDA derive synthetic variables that specifically maximize the between-class object dispersion. Because each discriminant function is a weighted linear combination of the measured predictor variables, the weights (called discriminant coefficients) can be used to define the contribution (i.e. importance) of each predictor variable to the observed discrimination between classes of objects. Results of the DFA can be visualized through a scatter plot with discriminant functions serving as synthetic axes.

Like PCA, DFA is based on the noniterative, eigenvector-based solution. As a result, it is more computationally efficient than iterative methods, and is appropriate for extremely large data sets. A model of class prediction can also be generated during discriminant analysis. Such model can be used to provide predictions for a new ('unknown') object based on the values of measured variables in that object (Putnam *et al.* 2013). Examples of the use of different DFA variants in ecological studies are available in the reports by Gilbert *et al.* (2012) and Koenig *et al.* (2011).

Orthogonal projections to latent structures discriminant analysis

Orthogonal projections to latent structures discriminant analysis (OPLS-DA) is based on the principle of partial least-squares (PLS) regression; PLS itself is an extension of the multiple linear regression model (Wold 1966). The goal of PLS regression is to predict response variable(s) Y from a (large) set of predictor variables X. PLS regression reduces the set of predictor variables to a smaller set of uncorrelated components and then performs least-squares regression on these components. In the process, both variables X and Y are projected to a new space. This process is called projection to latent (hidden) structures (Abdi & Williams 2010). Compared to multivariate regression, PLS has fewer assumptions (it can use predictor variables that are collinear and not independent; Tobias 1995). An additional modification to the PLS regression was developed by Trygg & Wold (2002), who introduced a way to remove systematic variation from the predictor variable data set X that is not correlated to the response variable data set Y, that is to remove variation within X that is orthogonal to Y. The advantage of such orthogonal projections to latent structures (OPLS) method is that a single latent variable (designated 'T') is used as a predictor of Y. All variability in X is separated into predictive (T) and uncorrelated information ($T_{\text{orthogonal}}$), and the two components can be analysed separately. In OPLS discriminant analysis, the Y is a binary class-designating variable, and the analysis aims to find the best separation between classes of objects along the T axis, while all variation unrelated to class separation is distributed along the $T_{\text{orthogonal}}$ axes (Westerhuis *et al.* 2010). Similar to DFA, each variable's loading indicates its contribution to the OPLS model. The generated model can then be applied to a new object to predict its class given the values of its X variables. The OPLS-DA output provides measures of the model fit (R^2), model predictive power (Q^2) and model accuracy based on a cross-validation procedure. Examples of OPLS-DA use in multivariate ecological analyses can be found in Shankar *et al.* (2013) and Ramadan *et al.* (2014).

Support vector machine

Support vector machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression (Vapnik 1979; Cortes & Vapnik 1995). Using a training set of objects separated into classes, the SVM algorithm finds a hyperplane in the data space that produces the largest minimum distance (called margin) between the objects that belong to different classes. The larger the margin, the lower the expected prediction (i.e. misclassification) error. As shown in Fig. 3, in comparison with DFA, which uses differences in class means to separate objects (Welling 2005), SVM only considers the objects on the edges of the margin (these are called support vectors). This increases class separation and reduces expected prediction error. SVM can be used for both linear and nonlinear-based discriminatory analyses (Gu *et al.* 2010), and is considered to be a superior DA method for high-dimensionality data sets with small sample size, especially when variable selection and filtering are included in the algorithm (Gokcen & Peng 2002; Knights *et al.* 2011). While SVM has yet to be used extensively in microbial ecology, at least one example is available (Yang *et al.* 2006).

Random forest

Random forest (RF) is an ensemble learning approach based on the use of decision (classification) trees. Decision tree learning seeks to construct a statistical model to predict the values of response variable(s) based on the given values of predictor variables. The model is obtained by iteratively partitioning the space of predictor variables and establishing a value of the response variable within each partition (Loh 2011). The result of such partitioning can be represented as a decision tree containing a set of if-then logical conditions. RF 'grows' many different decision trees for the same data set and thus is a type of ensemble classifier. The existing data set is used to build predictive classification trees and 'train' the algorithms. To classify a new object, the input data (values for all predictor variables) are provided to each tree, which then generates an output classification (called a vote). The RF then chooses the classification that garnered the most votes among all the trees in the forest.

Random forest has been shown to achieve very high discriminating power and accuracy of classification (Cutler *et al.* 2007). This is because RF voting consolidates decisions across thousands of individual trees that together provide very high overall classification accuracy despite potentially low classification accuracy of some individual trees. This technique also does not suffer from high variance or bias that single classification model can have. The accuracy of the constructed model can be calculated by the cross-validation technique (Japkowicz & Shah 2014), while the similarities among objects can be assessed by computing the average proximity (fraction of trees where objects occupy the same position on a particular tree) for each pair of objects. Variable importance in the classification model can be estimated in RF by permutation tests (Box 6). Specifically, the values of one of the predictor variables are randomly swapped in the data set and the number of votes for the correct assignment in all trees is then calculated. The difference in the number of votes between the 'untouched' forest and the variable-permuted forest represents the importance score of that variable in the RF model. Results of RF analysis can be visualized as a MDS scatter plot of matrix of proximities among objects. RF has been used in several recent studies as described by Yatsunenko *et al.* (2012), Lozupone *et al.* (2013) and Shankar *et al.* (2013).

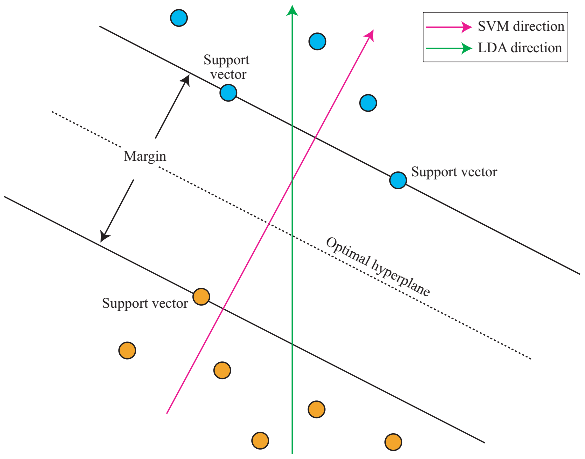


Fig. 3 Directions of class discrimination in linear and support vector machine-based discriminant analyses. Binary classification problem is depicted. Coloured circles represent different objects separated into two classes. The figure was adapted with permission from fig. 1 provided in the Gu *et al.* (2010) study.

In a recently published article, Knights *et al.* (2011) compared the performance of several different discriminatory and classification techniques. Data describing the community composition of human-associated microbiota were used as benchmarks, and several different algorithms including SVM and RF were compared. RF achieved the best performance among the compared techniques, followed closely by SVM applied to a filtered set of OTUs (Knights *et al.* 2011).

Box 6. Resampling tests of statistical significance

In many cases, after running a particular multivariate statistical analysis, an investigator desires to obtain some measure of confidence in the observed relationship or object separation. Such measure of confidence is usually acquired through a statistical test for significance. In parametric methods, the output is compared to the probability distribution model that was used as a basis for a particular technique. The probability of observing the obtained output by chance is then represented by a *P*-value. While this approach works well for methods such as regression and correlation analyses, for many ordination techniques no expected data distribution is defined before running the analyses. Thus, parametric statistical testing cannot be applied to these techniques.

For such cases, one approach to assess statistical confidence of the obtained output is through the use of resampling methods that include bootstrap, jackknife and permutation tests (Tukey 1958; Efron 1979; Desu & Raghavarao 2003). The principle of these techniques relies on the random resampling of the actual measured data to generate a reference distribution, which the observed distribution is tested against. In bootstrap tests, a subset of data is drawn randomly with replacement from a set of data points; in jackknife tests, a subset of data points is left out systematically and the analysis output remeasured; in permutation tests, object labels are randomly exchanged. The statistical significance (i.e. probability of observing the calculated output by chance) is then computed as the fraction of all simulated outcomes that are at least as extreme as the one originally observed. Because these tests make no assumptions about data distribution, they can be used practically on any data set irrespective of the assumed relationship among variables.

For example, we have incorporated the permutation analysis of interclass distance based on the Davies–Bouldin (DB) cluster separation index (Davies & Bouldin 1979) to assess the statistical significance of the observed object separation between classes in PCA (Shankar *et al.* 2013, 2015). Specifically, we calculated the DB distance metric as the ratio of the average interclass distance among all objects to the average of all intraclass distances. The permutation algorithm then randomly reassigned the object labels, and the DB distance metric was again computed. This process was repeated 10 000 times to produce a reference distribution of DB metric. The significance of object class-based separation observed in PCA ordination was calculated as the fraction of all permuted DB metric values that were equal or higher than the one obtained for the original data set (Shankar *et al.* 2013).

For constrained analyses, the significance of the constrained ordination model can be obtained by first calculating the fraction of total variance in the data set of response variables that is explained by canonical axes, and then comparing this statistic to the distribution of such fractions obtained through a permutation analysis of the data set (Šmilauer & Lepš 2014).

Examples of the use of multivariate techniques in microbial ecology research

Study example 1: PCoA

Koren *et al.* (2013) assessed the relative composition of human faecal microbial communities by exploratory ordination. The authors performed PCoA using Jensen–Shannon divergence metric calculated on high-throughput 16S rRNA gene sequence data obtained from the set of Human Microbiome Project faecal samples. The position of samples in the PCoA ordination space was found to reflect the relative abundances of key members of faecal microbial communities. To illustrate this concept, the authors visualized on a PCoA plot the relative abundances of different genera in each faecal sample according to a colour gradient as shown in Fig. 4. Three prominent genera of human gut microbiota—*Bacteroides, Prevotella* and *Clostridium*—revealed the most striking abundance gradients. Specifically, distribution of samples across PC1 axis reflected samples' *Bacteroides* relative abundances; sample distribution along PC2 axis corresponded to samples' *Prevotella* abundances. A similar strong gradient was observed for genus *Clostridium*; however, this effect contributed to the sample dispersion observed along both PC1 and PC2 axes (Fig. 4). Thus, the exploratory PCoA revealed that the abundances of *Bacteroides* and *Prevotella* served as the main drivers of variation in the faecal microbiota composition, concordant with their role as keystone genera of gut microbiota enterotypes (Arumugam *et al.* 2011).

Study example 2: CCA

The utility of CCA was demonstrated by ter Braak and Verdonschot, who applied CCA to the macroinvertebrate distribution data from two man-made stream tributaries (ter Braak & Verdonschot 1995). The goal was to identify major environmental gradients and their effects on the species distribution patterns. The data set consisted of counts data for species collected from each site as well as measurements of several environmental

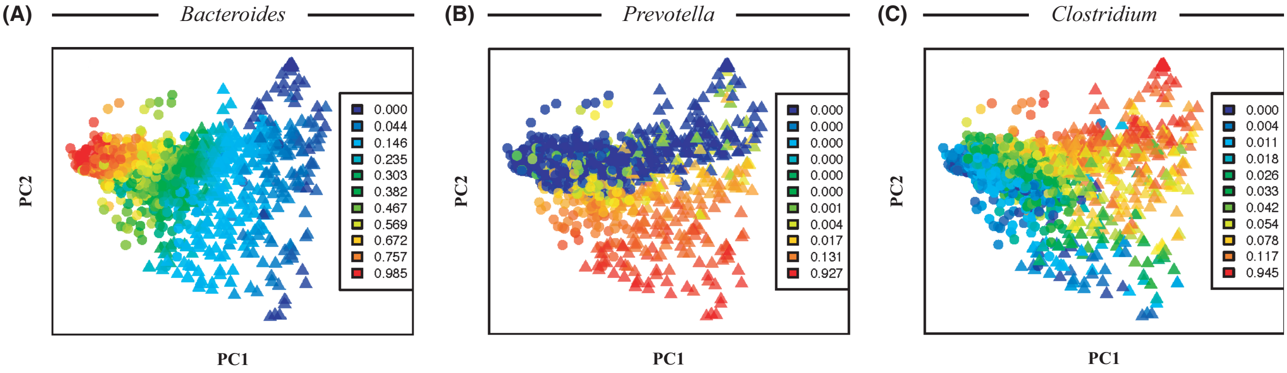


Fig. 4 Application of principal coordinates analysis (PCoA) to arrange faecal human samples based on microbiota structure. Jensen–Shannon divergence metric was used to calculate distances among all faecal samples obtained in the Human Microbiome Project (Peterson *et al.* 2009). Associations of sample distribution in PCoA space with relative abundances of genera *Bacteroides* (panel A), *Prevotella* (panel B) and *Clostridium* (panel C) are shown. Each sample point is colour-coded based on the abundance of the corresponding genus according to the colour gradient as shown in each figure legend. Samples are represented as either circles or triangles to denote microbiota enterotype cluster. The figure was adapted with permission from the supplementary figure S13 provided in the Koren *et al.* (2013) study.

variables. While this example does not describe a microbial system, it offers an excellent illustration of the interpretation of gradient analysis plots. The following insights can be made from the CCA triplot (Fig. 5):

1 The distance between site and species position on the triplot is indicative of the abundance of that species at that site: ‘the inferred abundance of a species is maximal if the site point coincides with the species point and decreases in all directions the farther away the site point is’ (ter Braak & Verdonschot 1995). Thus, we can conclude that species Et is abundant in sites U15–U19, whereas sites L15–L20 are optimal for Ms, Ss and Hd species (see Fig. 5 legend for species abbreviations).

2 Interpretation of quantitative environmental variables can be summarized as ‘the arrow points in the direction of maximum change in the value of the associated variable, and the arrow length is proportional to this maximum rate of change’ (ter Braak & Verdonschot 1995). Thus, the ‘source distance’ variable designates the distance of each site from the stream source, and the site ranking is consistent with the direction of the arrow. The difference in species distribution between L and U sites is small near the stream source (low ranks) but increases progressively the farther from the source the sites are (higher site ranks).

3 The relative positioning of environmental variables and species triangles indicates the optimum of species distribution along these environmental gradients. For example, for species Dl and Et, the distribution optimum is located the farthest along the gradient of electric conductivity. A similar interpretation can be extended to the qualitative variables. As stated by the authors, ‘sites that belong to a particular class’ (‘shrub’ or ‘no shrub’) ‘are scattered around the class point, simply because, by definition, each class is at the centroid of the sites that it contains’. By this interpretation, sites L15–L20 have high shrub coverage.

This example demonstrates how CCA triplots can provide a large number of insights and valuable interpretations, and such methods of direct gradient analyses are well suited to examine the relationships between environmental gradients and response variables (e.g. species).

Study example 3: PRC

Principal response curves analysis was used by Fuentes *et al.* (2014) to depict the time-dependent changes in distal gut microbiota composition of patients with *Clostridium difficile* infection (CDI) who have undergone FMT. The HITChip phylogenetic microarray was used to determine the distal gut microbiota composition from faeces collected from CDI patients before (day 0) and after FMT (days 14–70) and from their corresponding donors. PCA was first performed on the log-transformed microbiota data set (probe signal intensities) to visualize main gradients of variability (Fig. 6A). PC1 clearly separated all patient pre-FMT samples from donor and post-FMT samples. However, the samples of the latter group were all clustered together, which did not allow the assessment of the effect of factors such as time, interpersonal differences and donor variability on the distribution of samples in PCA space. To assess such possible effects more specifically, PRC analysis was used to isolate the

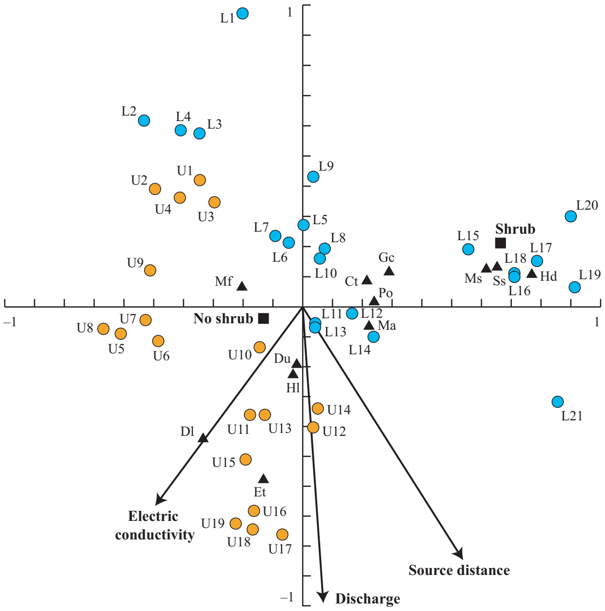


Fig. 5 Interpretation of canonical correspondence analysis (CCA) to relate macroinvertebrate species distribution to environmental sites and gradients. The figure displays a CCA triplot showing the relationships among sites (circles), species (triangles) and environmental variables (arrows and squares). Site labels (U and L) and colours (orange and cyan) represent the stream designations; site number denotes the ranked distance from the stream source. Black arrows denote quantitative environmental variables; black squares represent qualitative environmental variables. Black triangles indicate distribution optimum for each species. Select species are displayed on the plot and are abbreviated as follows: Ct—*Ceratopogonidae*, Dl—*Dendrocoelum lacteum*, Et—*Erpobdella testacea*, Gc—*Glossiphonia complanata*, Hl—*Haliplus lineatocollis*, Hd—*Helodidae*, Ma—*Micropsectra atrofasciata*, Mf—*Micropsectra fusca*, Ms—*Micropterna sequax*, Po—*Prodiamesa olivacea* and Ss—*Stictochironomus* sp. The figure was adapted with permission from fig. 3 provided in ter Braak & Verdonschot (1995) study.

relationships between a single environmental variable (time) and the observed microbiota changes (Fig. 6B). Assessment of the PRC chart indicated that gut microbiota changes occurred quickly after FMT procedure and community maintained similarity to that of the donor for at least 70 days, confirming a previously reported observation (Shankar *et al*. 2014). Additionally, the authors utilized the weights assigned by PRC analysis to different microbial groups to determine the main drivers of the changes in the microbiota profiles between pre- and post-FMT time points. Groups with positive weights such as Bacilli and Proteobacteria were shown to be higher in pre-FMT samples and were depleted as a function of time in the post-FMT samples, while groups with negative weights such as Bacteroidetes and *Clostridium* clusters IV and XIVa followed the opposite profile.

This study presented an excellent example of the utility of PRC to visualize time-dependent changes in community organization and showcased how partial ordination analysis can facilitate the interpretation of ordination results of complex data sets.

Study example 4: RF

Random forest discriminatory analysis was employed by Lozupone and colleagues to assess the distinctiveness of faecal microbiota between healthy subjects and HIV-infected patients (Lozupone *et al.* 2013). The authors utilized high-throughput sequencing of 16S rRNA gene to profile faecal microbiota from healthy controls and patients with chronic untreated HIV. The sequencing data were clustered to generate operational taxonomic unit (OTU) tables and subsequently subjected to RF analysis. RF generated a class prediction model (healthy vs. HIV), which was then used to classify unknown samples. Classification of 'unknown' samples achieved a 96% accuracy rate. This indicated that HIV-associated faecal microbiota displayed specific differences between healthy and HIV gut. Taking advantage of RF's ability to assign an importance score to each OTU variable, the authors identified discriminatory OTUs that marked sample donors as healthy or HIV. OTUs from genus *Bacteroides* were found to be depleted in HIV-infected patients compared to healthy controls. Conversely, members of *Erysipelotrichaceae*, *Veillonellaceae* and *Ruminococcoceae* were enriched in HIV-infected patients. In summary, RF analysis generated a model that used faecal microbiota OTU abundances to distinguish between healthy and HIV-infected subjects.

Limitations of multivariate analyses

Despite their undoubted usefulness, multivariate analyses also exhibit some limitations. The outputs of these algorithms are more difficult to interpret compared to those generated by univariate analyses (although it is much less time-consuming to interpret single multivariate analysis then hundreds of individual univariate outputs), and the assumptions of a particular method are sometimes not easy to assess or meet. It is also important to remember that revealed associations among variables and patterns of object distribution do not inherently imply causality (unless explanatory variables are experimental treatments), and that synthetic ordination axes or cluster groups might not necessarily match any biological effects or gradients. Finally, many multivariate techniques are very computationally demanding and require significant computing resources when applied to very large data sets.

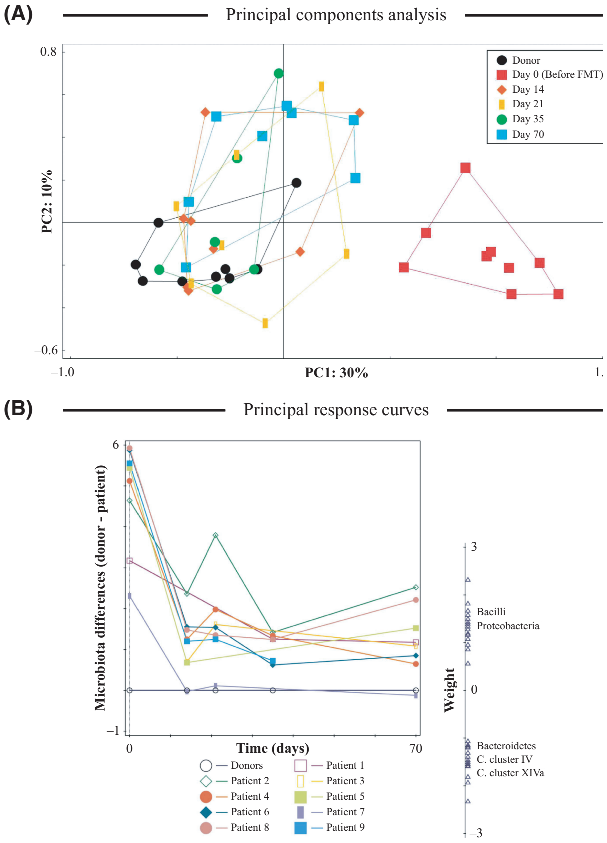


Fig. 6 Application of principal components and principal response curves analyses to visualize changes in microbial compositions over time. (Panel A) Principal components analysis (PCA) was performed on log-transformed HITChip microarray data to assess microbiota composition similarity between pre- (day 0) and post-faecal microbiota transplantation (post-FMT; days 14–70) faecal samples from *Clostridium difficile*-infected patients and faecal samples from FMT donors. (Panel B) Principal response curves analysis performed on the same data set. *X* axis shows collection time points. Microbial community changes over time are depicted on *Y* axis as the difference between microbiota composition of a given patient sample and that of the corresponding donor. Variable weights for individual microbial groups are depicted on the weight scale. The figure was adapted with permission from fig. 6 provided in the study by Fuentes *et al.* (2014).

The majority of multivariate statistical techniques discussed here define similarities and/or distances among objects based on the distribution of variables' values across these objects. The values of different variables are combined to generate the object distance or dissimilarity matrix. How different variables are combined impacts the output and interpretation of multivariate statistical techniques and can generate artefacts as described in detail by Warton *et al.* (2012). The authors point out that in real data sets different variables (taxon abundance, metabolite concentration, etc.) usually display unequal mean–variance relationships. This inequality can lead to the ordination displaying dispersion effects (differences among object classes in variables' variances) instead of the location effects (differences among object classes in variable means), because most distance-based techniques are not able to distinguish between these two effects (Fig. 7). The main limitation is not the ordination technique itself, but the distance metric used, and the authors show that most currently used distance metrics do not account correctly for the mean–variance relationship when combining data across variables. Note that the location-vs.-dispersion effects are less prominent in the constrained ordination, because constraints specifically focus on location differences in the data set (ter Braak & Šmilauer 2015).

Well-chosen data transformation applied to the data prior to statistical analysis (discussed in Box 2) can lessen the extent of mean–variance bias in multivariate data sets, but it does not remove it completely. Standardization of variable variances prior to multivariate analysis is not very helpful, because it reduces the effect of differences in variable means among classes and thus reduces true biological differences and statistical power of the analysis. Two solutions to the issue of unequal mean–variance distributions were offered by the authors. Standardization of within-class variance for each variable during the analysis solves the mean–variance issue but is very computationally intensive especially for techniques that use permutation to derive statistical significance of class separation (Box 6). Alternatively, multivariate GLM (Warton 2011) allows the choice of different mean–variance functions and was shown to perform well (O'Hara & Kotze 2010; Warton *et al.* 2012). To avoid possible artefacts in the multivariate analysis output, we recommend that the within-class variances for each variable are compared among all object classes, and are adjusted if found to be unequal.

Choice of the multivariate analysis

The multitude of different multivariate analytical techniques that were described above provides many choices for the selection of an algorithm to analyse a particular large-scale data set. In Fig. 8, we offer suggestions of some of the appropriate choices based on the research goal, data input structure and expected relationships among variables. Exploratory methods are in general applied first to extract the main gradients of variation in the data with unconstrained ordination techniques, and to identify groups of similar variables and/or objects with CLAs. To relate variability observed in the response variables to the values of additional explanatory variables, constrained gradient analysis should be performed. If the goal is to

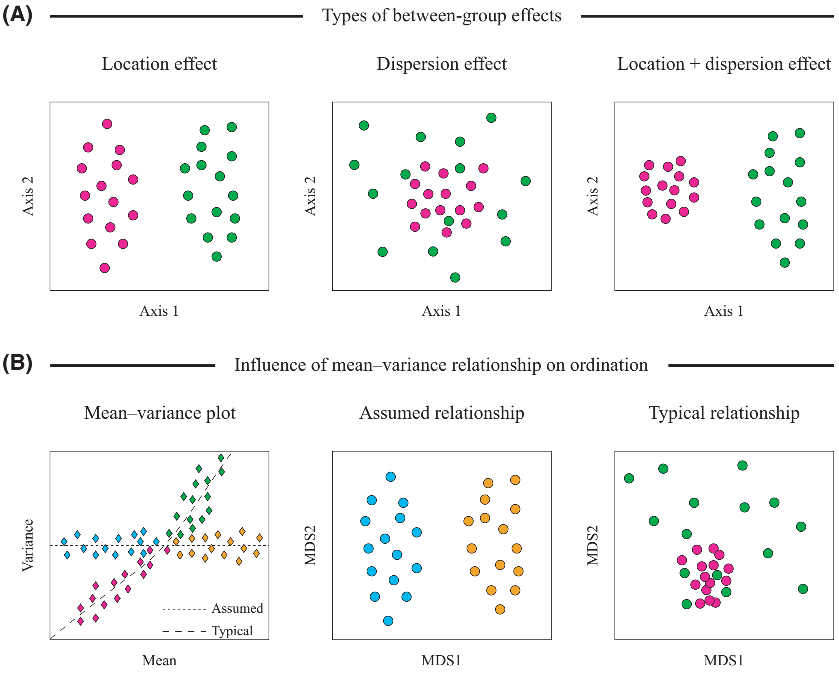


Fig. 7 Influence of mean–variance bias on the separation of location and dispersion effects in ordination analyses. Panels in section A show schematic representation of ordination results of object separation that display location, dispersion, or location and dispersion effects. Panels in section B display a schematic presentation of the effect of different mean–variance relationships on the object distribution in Euclidean distance-based multidimensional scaling ordination. Assumed vs. typical mean–variance relationships are shown on the left panel; diamonds represent different variables. Distribution of variables along the 'mean' axis represents the described difference (location effect) in the values of variables between classes of objects. Middle panel shows the visualization of ordination results when the equal variance assumption is correct. The right panel shows the ordination results for a more typically observed mean–variance relationship. Coloured circles represent different objects separated into two classes. The figure is based on figs 2 and 3a in Warton *et al.* (2012).

differentiate among a priori known classes of objects based on variable scores and/or to predict class membership of new/future observations, discriminant analyses and decision trees can be used.

The choice of a particular technique within a category usually depends on the input data structure and the assumed relationship among variables as we describe in Fig. 8. While available evidence based on the analysis of simulated data sets with known gradients does not provide a clear consensus on the best approach among linear, unimodal and distribution-free techniques (Austin 2013), several suggestions can be made.

In general, to analyse short environmental gradients and continuous variables expressed as absolute values, linear methods such as PCA and RDA are used most often (see figure in Box 5). To model long environmental gradients as well as relative abundances and count data, the application of unimodal ordination techniques such as (D)CA and CCA is usually advised (Austin 2007; Ramette 2007). Log-ratio transformation applied to the raw data can address the problem with data compositionality (Box 3). For data sets not conforming to most common statistical assumptions or distribution models, the use of nonparametric (distribution-free) methods is appropriate. When similarities among objects are desired to be assessed based on one of the ecological distance calculations, the use of PCoA, NMDS and db-RDA techniques is advocated.

When several different ordination outputs are available for the same set of objects, co-inertia and PAs can be used to assess the congruency of the results. To determine the fraction of the total variation in measured (response) variables attributed specifically to different environmental gradients, variation partitioning can be carried out (Box 5).

To find which variables are responsible for the observed differences among classes of objects, discriminatory analyses can be employed. OPLS-DA, SVM and RF all showed good performance in different comparative studies, with RF currently being favoured in microbial community comparisons (Knights *et al.* 2011). To avoid the problem of model overfitting, model regularization and variable selection and filtering should be applied during discriminatory analyses, especially for data sets with large number of variables (e.g. OTU tables) (Chen *et al.* 2015).

Closing comments

The development and continuing advances in the high-throughput molecular approaches have led to a dramatic shift in the type and amount of data generated by microbial researchers. It is a less common case now where experimental results can be analysed manually or with the use of simple parametric statistics and univariate analyses. Instead, the experimental data are often represented by large matrices of data points. This large-scale structure of obtained data limits researcher's ability to interpret the results without significant reduction in data set complexity. Multivariate analyses described in this review are well suited for this task,

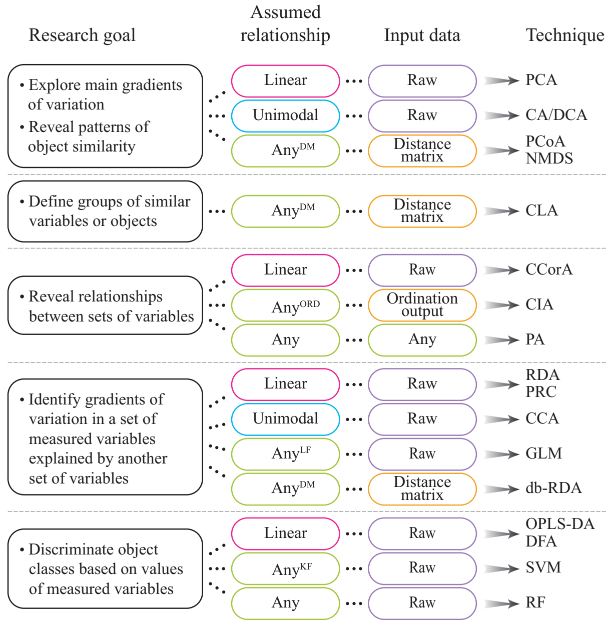

Research goal
Assumed relationship
Input data
Technique
• Explore main gradients of variation
• Reveal patterns of object similarity
Linear
Raw
PCA
Unimodal
Raw
CA/DCA
Any^DM
Distance matrix
PCoA
NMDS
• Define groups of similar variables or objects
Any^DM
Distance matrix
CLA
Linear
Raw
CCorA
• Reveal relationships between sets of variables
Any^ORD
Ordination output
CIA
Any
Any
PA
Linear
Raw
RDA
PRC
• Identify gradients of variation in a set of measured variables explained by another set of variables
Unimodal
Raw
CCA
Any^LF
Raw
GLM
Any^DM
Distance matrix
db-RDA
Linear
Raw
OPLS-DA
DFA
• Discriminate object classes based on values of measured variables
Any^KF
Raw
SVM
Any
Raw
RF

Fig. 8 Diagram of potential choices of multivariate techniques based on the research goal, assumed relationship among variables and input data structure. DM—assumed relationship depends on the chosen distance metric; ORD—assumed relationship depends on the ordination technique used; LF—assumed relationship depends on the chosen link function; KF—support vector machine model can be linear or nonlinear if a nonlinear kernel function is used.

and we provided a general overview of the most frequently used multivariate statistical methods to assist ecologists and other researchers in the choice of appropriate approaches for the analyses of their data sets.

The popularity of multivariate analyses continues to increase as more and more researchers start to take advantage of high-throughput experimental platforms. While the initial high-throughput studies tended to employ simpler exploratory techniques such as PCA and PCoA, there is a recent trend to move to interpretive and discriminatory ordination approaches and to perform more rigorous multivariate hypothesis testing. The use of more recently developed techniques including principal curves and surfaces (Hastie & Stuetzle 1989; De'ath 1999), dynamic factor analysis (Zuur *et al.* 2003), random-effect ordination (Walker & Jackson 2011), coreferentiality (Fesel 2012), multidimensional fuzzy set ordination (Roberts 2009) and fuzzy clustering (Bezdek 1981), elastic net regression (Zou & Hastie 2005), and regularized discriminant analyses (Friedman 1989; Zhao & Wong 2014; Chen *et al.* 2015) is expected to further advance these studies and improve the validity and robustness of the outcomes. All these techniques provide exciting opportunities to link ecological and functional measures of microbial communities with environmental gradients, host (patient) information, and time and space variables. Ecological theories can be tested at the whole community level, and microbial community structure and function can now be used to facilitate discrimination between different sites, environments, or between human health and disease (Lozupone *et al.* 2013; Shankar *et al.* 2015). All of such studies provide deeper appreciation for the importance of ecological analyses of biological systems, and will reveal further mechanisms of interactions among members of complex communities and between host and microbial partners.

Acknowledgements

We are grateful to Gengxin Li, Harry Khamis, Michael Raymer and anonymous reviewers for valuable comments on the manuscript. The work in authors' laboratory has been supported by the National Institutes of Health grants AT003423 and HD065575, by National Science Foundation grant DBI-1335772, and by Dayton Area Graduate Studies Institute.

References

Abdi H, Williams LJ (2010) Partial least square regression, projection on latent structure regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 433–459.

Aitchison J (1986) *The Statistical Analysis of Compositional Data*. Chapman & Hall, New York.

Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**, 32–46.

Anderson MJ, Willis TJ (2003) Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology*, **84**, 511–525.

Arthur D, Vassilvitskii S (2007) K-means plus plus: the advantages of careful seeding. *Proceedings of the Eighteenth Annual Acm-Siam Symposium on Discrete Algorithms*, 1027–1035.

Arumugam M, Raes J, Pelletier E *et al.* (2011) Enterotypes of the human gut microbiome. *Nature*, **473**, 174–180.

Austin M (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1–19.

Austin MP (2013) Inconsistencies between theory and methodology: a recurrent problem in ordination studies. *Journal of Vegetation Science*, **24**, 251–268.

Bady P, Doledec S, Dumont B, Fruget JF (2004) Multiple co-inertia analysis: a tool for assessing synchrony in the temporal variability of aquatic communities. *Comptes Rendus Biologies*, **327**, 29–36.

Ballabio D, Consonni V (2013) Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, **5**, 3790–3798.

Bassett SA, Young W, Barnett MP *et al.* (2015) Changes in composition of caecal microbiota associated with increased colon inflammation in interleukin-10 gene-deficient mice inoculated with enterococcus species. *Nutrients*, **7**, 1798–1816.

Bezdek JC (1981) *Pattern Recognition With Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell.

Borcard D, Legendre P, Drapeau P (1992) Partialling out the spatial component of ecological variation. *Ecology*, **73**, 1045–1055.

Borcard D, Gillet F, Legendre P (2011) *Numerical Ecology with R*. Springer, New York.

ter Braak CJF (1986) Canonical correspondence analysis—a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–1179.

ter Braak CJF (1987) *Unimodal Models to Relate Species to Environment*. Agricultural Mathematics Group, Wageningen.

ter Braak CJ (1988) Partial canonical correspondence analysis. In: *Classification Methods and Related Methods of Data Analysis* (ed. Bock HH), pp. 551–558. North-Holland, Amsterdam.

ter Braak CJF, Looman CWN (1986) Weighted averaging, logistic regression and the gaussian response model. *Vegetatio*, **65**, 3–11.

ter Braak CJF, Prentice IC (1988) A theory of gradient analysis. *Advances in Ecological Research*, **18**, 271–317.

ter Braak CJF, Šmilauer P (2015) Topics in constrained and unconstrained ordination. *Plant Ecology*, **216**, 683–696.

ter Braak CJF, Verdonschot PFM (1995) Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences*, **57**, 255–289.

Bray JR, Curtis JT (1957) An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, **27**, 326–349.

van den Brink PJ, ter Braak CJF (1999) Principal response curves: analysis of time-dependent multivariate responses of biological community to stress. *Environmental Toxicology and Chemistry*, **18**, 138–148.

van den Brink PJ, den Besten PJ, bij de Vaate A, ter Braak CJF (2009) Principal response curves technique for the analysis of multivariate biomonitoring time series. *Environmental Monitoring and Assessment*, **152**, 271–281.

Buttigieg PL, Ramette A (2015) A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiology Ecology*, **90**, 543–550.

Chen C, Zhang ZM, Ouyang ML *et al.* (2015) Shrunken centroids regularized discriminant analysis as a promising strategy for metabolomics data exploration. *Journal of Chemometrics*, **29**, 154–164.

Choulakian V (2001) Robust Q-mode principal component analysis in L 1. *Computational statistics & data analysis*, **37**, 135–150.

Claesson MJ, Jeffery IB, Conde S *et al.* (2012) Gut microbiota composition correlates with diet and health in the elderly. *Nature*, **488**, 178–184.

Clarke KR (1993) Nonparametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, **18**, 117–143.

Clarke KR, Gorley RN (2005) *PRIMER: Getting Started With v6*. PRIMER-E Ltd, Plymouth, UK.

Cortes C, Vapnik V (1995) Support-Vector Networks. *Machine Learning*, **20**, 273–297.

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Davies DL, Bouldin DW (1979) A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**, 224–227.

De'ath G (1999) Principal curves: a new technique for indirect and direct gradient analysis. *Ecology*, **80**, 2237–2253.

Desu MM, Raghavarao D (2003) *Nonparametric Statistical Methods for Complete and Censored Data*, 1st edn. Chapman and Hall/CRC, Boca Raton, Florida.

Doledec S, Chessel D (1994) Co-inertia analysis—an alternative method for studying species environment relationships. *Freshwater Biology*, **31**, 277–294.

Dray S, Chessel D, Thioulouse J (2003a) Co-inertia analysis and the linking of ecological data tables. *Ecology*, **84**, 3078–3089.

Dray S, Chessel D, Thioulouse J (2003b) Procrustean co-inertia analysis for the linking of multivariate datasets. *Ecoscience*, **10**, 110–119.

Driver HE, Kroeber AL (1932) *Quantitative Expression of Cultural Relationships*. University of California Press, Berkeley.

Dugard P, Todman J, Staines H (2014) *Approaching Multivariate Analysis. A Practical Introduction*. Routledge, New York.

Efron B (1979) Bootstrap methods—another look at the jackknife. *Annals of Statistics*, **7**, 1–26.

Eilam O, Zarecki R, Oberhardt M *et al.* (2014) Glycan degradation (GlyDeR) analysis predicts mammalian gut microbiota abundance and host diet-specific adaptations. *MBio*, **5**, e01526–14.

Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, **96**, 226–231.

Faust K, Sathirapongsasuti JF, Izard J *et al.* (2012) Microbial co-occurrence relationships in the human microbiome. *PLoS Computational Biology*, **8**, e1002606.

Ferreira L, Hitchcock DB (2009) A comparison of hierarchical methods for clustering functional data. *Communications in Statistics-Simulation and Computation*, **38**, 1925–1949.

Fesel C (2012) Coreferentiality: a new method for the hypothesis-based analysis of phenotypes characterized by multivariate data. *PLoS One*, **7**, e33990.

Fierer N, Leff JW, Adams BJ *et al.* (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences, USA*, **109**, 21390–21395.

Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Fox J (2008) *Applied Regression Analysis and Generalized Linear Models*. Sage Publications, Inc, London.

Friedman JH (1989) Regularized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165–175.

Friedman J, Alm EJ (2012) Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, **8**, e1002687.

Fuentes S, van Nood E, Tims S *et al.* (2014) Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent Clostridium difficile infection. *ISME Journal*, **8**, 1621–1633.

Gajer P, Brotman RM, Bai G *et al.* (2012) Temporal dynamics of the human vaginal microbiota. *Science Translational Medicine*, **4**, 132–152.

Gauch HG, Whittaker RH (1972) Coenocline simulation. *Ecology*, **53**, 446–451.

Gevers D, Kugathasan S, Denson LA *et al.* (2014) The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host & Microbe*, **15**, 382–392.

Gilbert JA, Steele JA, Caporaso JG *et al*. (2012) Defining seasonal marine microbial community dynamics. *ISME Journal*, **6**, 298–308.

Gokcen I, Peng J (2002) Comparing linear discriminant analysis and support vector machines. *Advances in Information Systems*, **2457**, 104–113.

Goodall DW (1954) Objective methods for the classification of vegetation. III. An essay on the use of factor analysis. *Australian Journal of Botany*, **1**, 39–63.

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Gower JC (1975) Generalized procrustes analysis. *Psychometrika*, **40**, 33–51.

Gu SC, Tan Y, He XG (2010) Discriminant analysis via support vectors. *Neurocomputing*, **73**, 1669–1675.

Guan XY, Wang JF, Zhao H *et al*. (2013) Soil bacterial communities shaped by geochemical factors and land use in a less-explored area, Tibetan Plateau. *BMC Genomics*, **14**, 820.

Hammer Ø, Harper DAT, Ryan PD (2001) PAST—Palaeontological statistics software package for education and data analysis. *Palaeontologia Electronica*, **4**, 1–9.

Hartmann M, Niklaus PA, Zimmermann S *et al*. (2014) Resistance and resilience of the forest soil microbiome to logging-associated compaction. *Isme Journal*, **8**, 226–244.

Hastie T, Stuetzle W (1989) Principal Curves. *Journal of the American Statistical Association*, **84**, 502–516.

Hastie T, Tibshirani R (1986) Generalized additive models. *Statistical Science*, **1**, 297–318.

Hastie T, Tibshirani R, Walther G (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, **63**, 411–423.

Heo MS, Gabriel KR (1998) A permutation test of association between configurations by means of the RV coefficient. *Communications in Statistics-Simulation and Computation*, **27**, 843–856.

Hildebrand F, Nguyen TLA, Brinkman B *et al*. (2013) Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biology*, **14**, R4.

Hill MO (1973) Reciprocal averaging—eigenvector method of ordination. *Journal of Ecology*, **61**, 237–244.

Hill MO (1974) Correspondence Analysis—neglected multivariate method. *Applied Statistics*, **23**, 340–354.

Hill MO (1979) *TWINSPAN: A FORTRAN Program for Arranging Multivariate Data in an Ordered Two-Way Table by Classification of the Individuals and Attributes*. Section of Ecology and Systematics, Cornell University, Ithaca, New York.

Hill MO, Gauch HG (1980) Detrended correspondence analysis—an improved ordination technique. *Vegetatio*, **42**, 47–58.

Hong PY, Lee BW, Aw M *et al*. (2010) Comparative Analysis of Fecal Microbiota in Infants with and without Eczema. *PLoS One*, **5**, e9964.

Hurley JR, Cattell RB (1962) The Procrustes program—producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, **7**, 258–262.

James G, Witten D, Hastie T, Tibshirani R (2014) *An Introduction to Statistical Modelling*. Springer, New York.

Japkowicz N, Shah M (2014) *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, Cambridge, UK.

Jones SE, Shade AL, McMahon KD, Kent AD (2007) Comparison of primer sets for use in automated ribosomal intergenic spacer analysis of aquatic bacterial communities: an ecological perspective. *Applied and Environmental Microbiology*, **73**, 659–662.

Knights D, Costello EK, Knight R (2011) Supervised classification of human microbiota. *Fems Microbiology Reviews*, **35**, 343–359.

Koenig JE, Spor A, Scalfone N *et al*. (2011) Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences, USA*, **108**(Suppl 1), 4578–4585.

Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biological cybernetics*, **43**, 59–69.

Koren O, Knights D, Gonzalez A *et al*. (2013) A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *Plos Computational Biology*, **9**, e1002863.

Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1–27.

Kucera M, Malmgren BA (1998) Logratio transformation of compositional data—a resolution of the constant sum constraint. *Marine Micropaleontology*, **34**, 117–120.

Legendre P, Anderson MJ (1999) Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, **69**, 1–24.

Legendre P, Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, **129**, 271–280.

Legendre P, Legendre L (2012) *Numerical Ecology*, 3rd edn. Elsevier, Amsterdam.

Lisboa FJG, Peres-Neto PR, Chaer GM *et al*. (2014) Much beyond Mantel: bringing Procrustes association metric to the plant and soil ecologist's toolbox. *PLoS One*, **9**, e101238.

Littman RA, Willis BL, Pfeffer C, Bourne DG (2009) Diversities of coral-associated bacteria differ with location, but not species, for three acroporid corals on the Great Barrier Reef. *FEMS Microbiology Ecology*, **68**, 152–163.

Liu Q (1997) Variation partitioning by partial redundancy analysis (RDA). *Environmetrics*, **8**, 75–85.

Lockyear K (2000) Experiments with detrended correspondence analysis. In: *Computer Applications and Quantitative Methods in Archaeology* (eds. Lockyear K, Sly TJT, Mihailescu-Birliba V), pp. 9–17. Archaeopress, Oxford.

Loh WY (2011) Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **1**, 14–23.

Lorenzo-Seva U, van de Velden M, Kiers HAL (2009) CAR: a MATLAB package to compute correspondence analysis with rotations. *Journal of Statistical Software*, **31**, 1–14.

Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bahler J (2015) Proportionality: a valid alternative to correlation for relative data. *PLOS One*, **11**, e1004075.

Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environment Microbiology*, **71**, 8228–8235.

Lozupone CA, Hamady M, Kelley ST, Knight R (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, **73**, 1576–1585.

Lozupone CA, Li M, Campbell TB *et al.* (2013) Alterations in the gut microbiota associated with HIV-1 infection. *Cell Host & Microbe*, **14**, 329–339.

Manly BFJ (2004) *Multivariate Statistical Methods: A Primer*, 3rd edn. Chapman and Hall, Boca Raton.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Mason OU, Scott NM, Gonzalez A *et al.* (2014) Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *ISME Journal*, **8**, 1464–1475.

McHardy IH, Goudarzi M, Tong M *et al.* (2013) Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome*, **1**, 17.

Minchin PR (1987) An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, **69**, 89–107.

Nelder JA, Wedderburn RWM (1972) Generalized Linear Models. *Journal of the Royal Statistical Society Series A-General*, **135**, 370–384.

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.

O'Hara RB, Kotze DJ (2010) Do not log-transform count data. *Methods in Ecology and Evolution*, **1**, 118–122.

Paliy O, Foy BD (2011) Mathematical modeling of 16S ribosomal DNA amplification reveals optimal conditions for the interrogation of complex microbial communities with phylogenetic microarrays. *Bioinformatics*, **27**, 2134–2140.

Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**, 559–572.

Perez-Cobas AE, Artacho A, Ott SJ *et al.* (2014) Structural and functional changes in the gut microbiota associated to Clostridium difficile infection. *Frontiers in Microbiology*, **5**, 335.

Peterson J, Garges S, Giovanni M *et al.* (2009) The NIH Human Microbiome Project. *Genome Research*, **19**, 2317–2323.

Prideaux L, Kang S, Wagner J *et al.* (2013) Impact of Ethnicity, Geography, and Disease on the Microbiota in Health and Inflammatory Bowel Disease. *Inflammatory Bowel Diseases*, **19**, 2906–2918.

Putnam RA, Mohaidat QI, Daabous A, Rehse SJ (2013) A comparison of multivariate analysis techniques and variable selection strategies in a laser-induced breakdown spectroscopy bacterial classification. *Spectrochimica Acta Part B: Atomic Spectroscopy*, **87**, 161–167.

Rajilic-Stojanovic M, Maathuis A, Heilig HG *et al.* (2010) Evaluating the microbial diversity of an in vitro model of the human large intestine by phylogenetic microarray analysis. *Microbiology*, **156**, 3270–3281.

Ramadan Z, Xu H, Laflamme D *et al.* (2014) Fecal microbiota of cats with naturally occurring chronic diarrhea assessed using 16S rRNA gene 454-pyrosequencing before and after dietary treatment. *Journal of Veterinary Internal Medicine*, **28**, 59–65.

Ramette A (2007) Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, **62**, 142–160.

Rao CR (1964) The use and interpretation of principal component analysis in applied research. *Sankhya A*, **26**, 329–358.

Ringel-Kulka T, Cheng J, Ringel Y *et al.* (2013) Intestinal microbiota in healthy U.S. young children and adults-a high throughput microarray analysis. *PLoS One*, **8**, e64315.

Roberts DW (2009) Comparison of multidimensional fuzzy set ordination with CCA and DB-RDA. *Ecology*, **90**, 2622–2634.

Rodgers JL, Nicewander WA, Toothaker L (1984) Linearly independent, orthogonal, and uncorrelated variables. *American Statistician*, **38**, 133–134.

Schnorr SL, Candela M, Rampelli S *et al.* (2014) Gut microbiome of the Hadza hunter-gatherers. *Nature Communications*, **5**, 3654.

Schwartz S, Friedberg I, Ivanov IV *et al.* (2012) A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome Biology*, **13**, r32.

Shakya M, Gottel N, Castro H *et al.* (2013) A multifactor analysis of fungal and bacterial community structure in the root microbiome of mature Populus deltoides trees. *PLoS One*, **8**, e76382.

Shankar V, Agans R, Holmes B, Raymer M, Paliy O (2013) Do gut microbial communities differ in pediatric IBS and health? *Gut Microbes*, **4**, 347–352.

Shankar V, Hamilton MJ, Khoruts A *et al.* (2014) Species and genus level resolution analysis of gut microbiota in Clostridium difficile patients following fecal microbiota transplantation. *Microbiome*, **2**, 13.

Shankar V, Homer D, Rigsbee L *et al.* (2015) The networks of human gut microbe-metabolite associations are different between health and irritable bowel syndrome. *ISME Journal*, doi:10.1038/ismej.2014.258.

Shevlyakov G, Smirnov P (2011) Robust estimation of the correlation coefficient: an attempt of survey. *Austrian Journal of Statistics*, **40**, 147–156.

Šmilauer P, Lepš J (2014) *Multivariate Analysis of Ecological Data Using* CANOCO *5*. Cambridge University Press, Cambridge.

Sridharan GV, Choi K, Klemashevich C *et al.* (2014) Prediction and quantification of bioactive microbiota metabolites in the mouse gut. *Nature Communications*, **5**, 5492.

Szekely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, **35**, 2769–2794.

Ten Berge JMF (1977) Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, **42**, 267–276.

Thureborn P, Lundin D, Plathan J *et al.* (2013) A metagenomics transect into the deepest point of the Baltic Sea reveals clear stratification of microbial functional capacities. *PLoS One*, **8**, e74983.

Tobias RD (1995) An introduction to partial least squares regression. *Proceedings of the 20th Annual SAS Users Group International Conference*, 2–5.

Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, **16**, 119–128.

Tukey JW (1958) Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics*, **29**, 614.

Ushio M, Yamasaki E, Takasu H *et al.* (2015) Microbial communities on flower surfaces act as signatures of pollinator visitation. *Scientific Reports*, **5**, 8695.

Van den Wollenberg AL (1977) Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, **42**, 207–219.

Vapnik V (1979) *Estimation of Dependences Based on Empirical Data*. Nauka, Moscow [in Russian].

Volis S, Dorman M, Blecher M, Sapir Y, Burdeniy L (2011) Variation partitioning in canonical ordination reveals no effect of soil but an effect of co-occurring species on translocation success in Iris atrofusca. *Journal of Applied Ecology*, **48**, 265–273.

Walker SC, Jackson DA (2011) Random-effects ordination: describing and predicting multivariate correlations and cooccurrences. *Ecological Monographs*, **81**, 635–663.

Wang X, Hu M, Xia Y, Wen X, Ding K (2012a) Pyrosequencing analysis of bacterial diversity in 14 wastewater treatment systems in China. *Applied and Environment Microbiology*, **78**, 7042–7047.

Wang XH, Eijkemans MJC, Wallinga J *et al.* (2012b) Multivariate approach for studying interactions between environmental variables and microbial communities. *PLoS One*, **7**, e50267.

Warton DI (2011) Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics*, **67**, 116–123.

Warton DI, Wright ST, Wang Y (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, **3**, 89–101.

Welling M (2005) Fisher linear discriminant analysis. *Department of Computer Science, University of Toronto*, **3**, 1–4.

Westerhuis JA, van Velzen EJJ, Hoefsloot HCJ, Smilde AK (2010) Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics*, **6**, 119–128.

Whittaker RH (1956) Vegetation of the great smoky mountains. *Ecological Monographs*, **26**, 1–69.

Whittaker RH (1967) Gradient analysis of vegetation. *Biological Reviews of the Cambridge Philosophical Society*, **42**, 207–264.

Whittaker RH, Levin SA, Root RB (1973) Niche, habitat, and ecotope. *American Naturalist*, **107**, 321–338.

Withman B, Gunasekera TS, Beesetty P, Agans R, Paliy O (2013) Transcriptional responses of uropathogenic Escherichia coli to increased environmental osmolality caused by salt or urea. *Infection and Immunity*, **81**, 80–89.

Wold H (1966) Estimation of principal components and related models by iterative least squares. In: *Multivariate Analysis*, (ed. Krishnaiah PR) pp. 391–420. Academic Press, New York.

Yan Q, Bi Y, Deng Y *et al.* (2015) Impacts of the Three Gorges Dam on microbial structure and potential function. *Scientific Reports*, **5**, 8605.

Yang CY, Mills D, Mathee K *et al.* (2006) An ecoinformatics tool for microbial community studies: supervised classification of Amplicon Length Heterogeneity (ALH) profiles of 16S rRNA. *Journal of Microbiological Methods*, **65**, 49–62.

Yatsunenko T, Rey FE, Manary MJ *et al.* (2012) Human gut microbiome viewed across age and geography. *Nature*, **486**, 222–227.

Yee TW (2004) A new technique for maximum-likelihood canonical Gaussian ordination. *Ecological Monographs*, **74**, 685–701.

Yelland PM (2010) An Introduction to Correspondence Analysis. *The Mathematica Journal*, **12**, 1–23.

Zhang C, Zhang M, Pang X *et al.* (2012a) Structural resilience of the gut microbiota in adult mice under high-fat dietary perturbations. *ISME Journal*, **6**, 1848–1857.

Zhang J, Peréz O, Lalles JP, Smidt H (2012b) Influence of antibiotic treatment of sows on intestinal microbiota in their offsprings. In: *Program and Book of Abstracts of the 12th International Symposium Digestive Physiology of Pigs*, p. 50.

Zhang Y, Lu Z, Liu S *et al.* (2013) Geochip-based analysis of microbial communities in alpine meadow soils in the Qinghai-Tibetan plateau. *BMC Microbiology*, **13**, 72.

Zhang C, Yin A, Li H *et al.* (2015) Dietary modulation of gut microbiota contributes to alleviation of both genetic and simple obesity in children. *EBioMedicine*, **2**, 968–984.

Zhao HT, Wong WK (2014) Regularized discriminant entropy analysis. *Pattern Recognition*, **47**, 806–819.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, **67**, 301–320.

Zuur AF, Tuck ID, Bailey N (2003) Dynamic factor analysis to estimate common trends in fisheries time series. *Canadian Journal of Fisheries and Aquatic Sciences*, **60**, 542–552.

Both authors were involved in the planning and writing of this review.

Data accessibility

New data were not generated for this review.