On selecting an appropriate multivariate analysis

N. C. Kenkel

Department of Botany, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2 (e-mail: kenkel@cc.umanitoba.ca). *Received 24 August 2005, accepted 15 March 2006.*

Kenkel, N. C. 2006. **On selecting an appropriate multivariate analysis.** Can. J. Plant Sci. **86**: 663–676. The broad objective of multivariate data analysis in biology is to summarize associations among species (the dependent or response variables), and to elucidate species responses to one or more environmental factors (the independent or predictor variables). This objective is achieved by reducing the dimensionality of variable space to an efficient, low-dimensional summative model of the underlying data structure that reflects the coordinated response of species to environmental factors. While multivariate methods have proven indispensable for analyzing both experimental and survey data in the biological sciences, considerable confusion persists regarding the selection of appropriate analytical strategies. The selection of an appropriate analytical strategy, which includes important decisions regarding data transformation, variable standardization and methodological approach, should be based on fundamental considerations of statistical appropriateness, data structure, and study objectives. Unfortunately, past and more recent assessments of multivariate analytical strategies have been based largely on empirical models of questionable relevance. This empirical approach has led to misleading recommendations and erroneous generalizations regarding the relative efficacy of the available multivariate methods. This paper dispels these misleading recommendations and provides some general guidelines for selecting appropriate data transformations, variable standardizations and methodological approaches in the multivariate analysis of biological data.

Key words: Ordination, canonical analysis, co-inertia analysis, principal component analysis, correspondence analysis, non-metric multidimensional scaling

Kenkel, N. C. 2006. **Choix d'une méthode convenable d'analyse à variables multiples**. Can. J. Plant Sci. **86**: 663–676. En biologie, l'analyse à variables multiples des données a pour objectif général de résumer les associations entre espèces (la variable dépendante ou la variable-réponse) et d'élucider la réaction de l'espèce à un ou à plusieurs paramètres environnementaux (les variables indépendantes ou explicatives). On y parvient en diminuant le nombre de dimensions de l'espace des variables jusqu'à obtenir un modèle efficace à peu de dimensions qui résumera la structure sous-jacente des données et illustrera la réaction coordonnée de l'espèce aux facteurs environnementaux. Bien que les méthodes à variables multiples s'avèrent indispensables à l'analyse des données expérimentales et des données relevées sur le terrain en biologie, une grande confusion règne dans le choix de la méthode appropriée. Pareil choix suppose en effet d'importantes décisions au niveau de la conversion des données, de l'uniformisation des variables et de la méthodologie. Il devrait donc reposer sur des considérations fondamentales, notamment la valeur statistique, l'organisation des données et les buts de l'exercice. Malheureusement, les évaluations anciennes et d'autres, plus récentes, des méthodes à variables multiples reposent dans une large mesure sur des modèles empiriques dont la pertinence peut être mise en doute. Cette approche empirique a débouché sur des recommandations fallacieuses et des généralisations erronées concernant l'efficacité relative des méthodes existantes. Cet article rejette ces recommandations et donne quelques lignes directrices générales qui faciliteront le choix d'une approche convenable à la transformation des données, à l'uniformisation des variables et à la méthodologie pour l'analyse à variables multiples des données biologiques.

Mots clés: Ordination, analyse canonique, analyse par co-inertie, analyse en composantes principales, analyse par correspondance, analyse multidimensionnelle non métrique

Classic statistical methods used in field biology were developed to address the specific needs of agriculturalists working with highly controlled experimental systems (Digby and Kempton 1987; Sokal and Rohlf 1995). Characteristic features of such controlled experiments include the manipulation of experimental factors over a relatively narrow range, the minimization of potentially confounding uncontrolled factors through careful site selection and appropriate experimental design, and emphasis on the analysis of a single response variable such as yield. These features allowed for the development of a statistical theory based on two important assumptions: that the factors affecting variable response patterns are additive, and that the residual response is normally distributed (Digby and Kempton 1987). Provided that these assumptions are met (exactly, or at least approximately), classic statistical theory provides powerful analytical methods for detecting departures from the null hypothesis (Mead 1988; Sokal and Rohlf 1995).

While classical statistical methods are well-suited to the analysis of data arising from highly controlled experiments, they are of limited utility for analyzing data arising from surveys or experiments undertaken in natural or semi-natural ecosystems (e.g., rangeland, agricultural weed communities). Consider for example the Park Grass Experiment,

Presented at the Plant Canada 2005 Symposium "Contemporary Issues in Statistics, Data Management, Plant Biology and Agriculture Research".

Abbreviations: **CA**, correspondence analysis; **CCor**, canonical correlation analysis; **COIA**, co-inertia analysis; **DTA**, detrended correspondence analysis; **NMDS**, non-metric multidimensional scaling; **PCA**, principal component analysis; **PCoA**, principal coordinate analysis; **RDA**, redundancy analysis

established at Rothamsted, England, in 1856 and used to investigate the long-term effects of nutrient manipulation on the productivity and biodiversity of semi-natural grasslands (Digby and Kempton 1987). This manipulative experiment differs from classic controlled experiments in a great many ways. Perhaps most importantly, the data are multivariate: responses of a large number of species are simultaneously recorded. There are also very wide ranges in species yields, and many species are entirely absent (zero yields) from most plots. Even after removing rare species from the Park Grass Experiment data, for example, over 50% of the data matrix contains zero values (Digby and Kempton 1987). Furthermore, the interactions of species responses and environmental factors, and those among species in the plots, are highly complex. This complexity necessarily leads to study objectives that are fundamentally different from simple estimation and hypothesis testing, the standard objectives of classic statistical methods. Specifically, the objectives of multivariate data analysis are to summarize associations among species, and to elucidate species responses to environmental factors (Legendre and Legendre 1998).

Multivariate statistical methods were first described in the early years of the 20th century, but computational challenges precluded their wide application until the advent of computers. They were introduced into the biological literature in the 1960s as effective methods for the analysis and display of complex data structures (Rao 1964; Seal 1964; Orloci 1966). Today multivariate statistical methods are indispensable analytical tools in the agricultural, biological and environmental sciences, as well as such diverse fields as psychology, economics and medicine. Yet despite nearly 40 years of application, and innumerable research articles discussing the relative efficacy and utility of various multivariate models, there persists a great deal of confusion regarding the selection and application of multivariate techniques. Most studies comparing multivariate methods and strategies are empirical and based on inductive reasoning. As a result, recommendations largely reflect the biases inherent in conclusions derived from empirical data, as well as the biases of individual researchers. While many researchers have maintained a certain objectivity in their assessments, a number of misleading and erroneous claims regarding the relative efficacy of multivariate methods persist in the literature. In some cases, unwarranted pronouncements have been made regarding the application of multivariate methods, which, in my opinion, have done a great disservice to the scientific community by perpetuating unwarranted generalizations. Unfortunately, a number of researchers and reviewers of scientific manuscripts have adopted these views, strongly advocating the use of a particular multivariate method to the exclusion of all others. Such a simplifying perspective fails to give due consideration to the issues most critical to the selection of an appropriate multivariate analytical strategy: statistical relevance, data structure, and study objectives.

It is against this current state of affairs that I have written the present article. My intent is not to present a statistical summary of multivariate methods – there are plenty of articles and monographs on that subject – but rather to provide some general guidelines for selecting appropriate multivariate analytical strategies based on fundamental considerations of statistical relevance, data structure, and study objectives. My objective is to dispel prevailing misconceptions regarding the relative efficacy of multivariate methods, which initially appeared in the ecological literature in the early 1970s and have been uncritically perpetuated even to this day.

MULTIVARIATE DATA STRUCTURES

Data are said to be multivariate (or multivariable) when a sample survey or experiment results in measurements of more than one variable in each sampling unit. The resultant data are typically represented as a matrix of attribute values, with P rows of variables and N columns of sampling units. An example is a biotic (species) data set consisting of abundance values of P_1 plant species (dependent or response variables) within each of N plots (sampling units) located at random within a grazed pasture. Possibly, but not necessarily, a second, abiotic (environment) data set consisting of values for P_2 soil nutrient measurements (independent or predictor variables) may be obtained from the same N plots, with the ultimate objective of predicting species composition from soil nutrient status (or vice versa). The defining feature of multivariate data analysis is that the P variables are considered simultaneously to investigate their coordinated response. It is this coordinated response that produces underlying trends and patterns in the data.

It is important to recognize and distinguish among different types of multivariate data structures, since data structure plays a critical role in determining the appropriate analytical strategy. The analytical strategy itself encompasses three related decisions: data transformation, variable standardization, and methodological approach. The following discussion focuses on three types of observed data structures: continuous abiotic (environmental) survey data, continuous biotic (species) survey data, and categorical contingency data. Simulated ecological data are also discussed, since they are widely used to assess the relative efficacy of multivariate analysis methods and strategies.

Observed Data Structures

Abiotic (Environmental) Survey Data

Examples of abiotic (environmental) survey data include soil nutrients measured in fields, and climatic factors measured at a set of discrete locations. Common features of such data include continuous predictor variables measured in different units (variables lacking a common scale), variable distributions that are positively skewed (log-normal distributions), and an absence of zero values. The appropriate analytical strategy for such data typically involves logarithmic transformation, variable standardization to z-scores, and the application of linear multivariate methods.

Biotic (Species) Survey Data

Examples of biotic (species) survey data include measured abundances of weed species in a series of fields, and insect species in a set of light traps. Common features of such data include continuous variables expressed in the same units (variables have a common scale, for example counts or related measures of absolute abundance), variable distribu-

tions that are highly positively skewed (log-linear distributions), and the presence of many zero values. The presence of zeros in such data is attributable to two factors: the distribution of species commonness and rarity, and niche limitations on the occurrence of species within some sampling units. The appropriate analytical strategy normally involves logarithmic transformation but no variable standardization (since measures are made on the same scale), and the application of linear or non-linear multivariate methods. Linear methods are appropriate when the zero values in the data largely reflect absences of rare species, whereas non-linear methods are appropriate when niche limitations are paramount (that is, when the sample encompasses a very broad range of environmental variation). For survey and experimental data in agriculture and related disciplines, linear multivariate models are generally appropriate since the sampled range of environmental variation is normally modest.

Contingency Data

Biotic contingency data most commonly arise through compilation of a series of biotic surveys or sub-samples from separate areas, and/or by combining variables and/or sampling units into broader categories (e.g., Greenacre and Vrba 1984). The resulting categorical data are in contingency form since the matrix elements represent counts (or abundances) of individuals in a cross-classification defined by the row and column categories. As an example, consider a study undertaken to survey the species composition/abundance of weed species in a series of 35 agricultural fields. Each field is first sampled using 20 randomly located plots in which weed species abundances are measured (i.e., a single biotic survey). The survey data are then compiled into a contingency table, by computing frequencies or means of weed species abundances (P row categories) in each field (N = 35 column categories). The variables may also be combined to form broader categories, for example by amalgamating taxonomic species into life-form classes (annual, biennial, perennial herb, perennial woody). Similarly, the 35 fields may also be classified into broader categories, for example productivity classes (low, medium, high and very high yielding fields).

While biotic survey data can be viewed as special cases of contingency data (see Legendre and Legendre 1998), it is important to distinguish these two data structures. The distinguishing features of biotic contingency data are a much larger sample size (they are often obtained by compiling numerous smaller biotic surveys, as in the example above), and much broader spatial (environmental) and/or temporal scales. Consider for example a study to determine variation in insect pollinators along a broad elevation (environmental) gradient, extending from sea level to the alpine. At a given elevation, a detailed biotic survey (e.g., 50 overnight light traps) would be required to enumerate fully the community of insect pollinators. Similar biotic surveys (i.e., 50 light traps each) would be undertaken at each elevation level. Finally, the individual biotic surveys would be compiled to form a contingency matrix (rows = insect species as variables, columns = elevation categories as sampling units). Note that biotic survey data at a given elevation level encompass a relatively narrow range of environmental variation, and could be separately analyzed using a linear multivariate method. By contrast, the contingency table summarizes variation over a much broader range of environmental variation, and therefore requires a non-linear (unimodal response) multivariate approach. The appropriate analytical strategy is to apply correspondence analysis on unstandardized, untransformed data.

As noted above, biotic survey data do not (and should not) encompass a wide range of environmental variation – no one would place great faith in the insect pollinator study outlined above if only a single overnight light trap was established at each elevation level (i.e., a biotic survey data set with insect species as variables, light traps as sampling units). While a clear distinction between biotic survey and contingency data is rarely made, such a distinction is critically important in determining the appropriate analytical strategy for multivariate analysis.

Simulated Data Structures

Assessments of multivariate techniques are typically based on the ability of methods to recover the underlying structure of simulated biotic data, an approach that dates back to the pioneering work of plant ecologists (Swan 1970; Gauch and Whittaker 1972a,b). This inductive approach was elaborated upon by Minchin (1987a,b) and others, and many contemporary researchers continue to use this approach for evaluation purposes (e.g., McCune and Grace 2002). A major advantage of using simulation modeling is that the underlying data structure is known, making the approach useful for comparing multivariate strategies and approaches under specified conditions (e.g., Bradfield and Kenkel 1987; Podani and Miklos 2002). If simulations are used to assess multivariate methods and strategies, however, then it must be convincingly demonstrated that simulated data are representative of data structures expected from the enumeration of natural systems (Minchin 1987a,b).

Simulated data structures are based on Whittaker's (1956) pioneering work on plant species response to one or two very broad environmental gradients, in a method known as direct gradient analysis (Whittaker 1978). In this approach, idealized species response curves along one or more complex environmental gradients are obtained by averaging or smoothing observed abundance data obtained from extensive field surveys. The data for common species are then smoothed to obtain idealized unimodal curves that represent realized niche responses along the gradient or gradients studied (Digby and Kempton 1987).

This idealized niche-based direct gradient model has been used by biologists for over 30 years to compare and evaluate multivariate methods and approaches. Simulated data are generated by representing species responses to a single environmental gradient (coenocline) as idealized unimodal Gaussian curves, or to two orthogonal environmental gradients (coenoplane) as unimodal Gaussian response surfaces (Gauch and Whittaker 1972a, 1976). While numerous modifications have been subsequently made to the basic model [e.g., introducing random variation (Minchin 1987a)], the basic and essential features of the gradient model have been retained to this day.

It is instructive to consider the nature of the simulated data arising from the coenoplane model:

1. Variations in species abundances are entirely attributable to species responses to one or two dominant environmental gradients.
2. Simulated species curves are idealized and smoothed (averaged) niche-based responses.
3. Only common species are considered.
4. Environmental gradients are orthogonal (uncorrelated), and all gradient combinations occur with equal frequency.

Are such data representative of natural systems, or more specifically of observed biotic data structures? Let us consider each point in turn:

1. In practice, species respond to a wide variety of factors, including numerous environmental factors (not just one or two) as well as disturbance, site history, dispersal limitations, and so forth.
2. Biotic survey and contingency data are notoriously "noisy", and a given species may not occur in the data where it "should" (i.e., according to a niche-based response curve). Thus, observed species responses, obtained through sampling, will differ substantially from idealized ones.
3. Biotic data are characterized by log-linear species-abundance distributions, resulting in many rare species that have important consequences on data structure.
4. Environmental variables are not independent, and some environmental combinations may be absent or rare, e.g., for the environmental gradients soil moisture and nutrient status, the combination of "very well-drained" and "nutrient-rich" is rarely encountered.

The inescapable conclusion is that the coenoplane model produces simulated data structures that are fundamentally different from those expected of biotic survey or contingency data. This calls into serious question the inductive approach of using simulated coenoplane data to evaluate the efficacy of multivariate methods and approaches.

OBJECTIVES OF MULTIVARIATE ANALYSIS

The objective of ordination is to reduce the dimensionality of a set of variables in order to obtain a low-dimensional summative model of the underlying multivariate data structure. Ordination methods achieve this goal by producing a statistically optimized arrangement of the sampling units along a reduced number of derived ordination axes. In addition, useful ordination methods produce weights and/or biplot scores for variables along these derived axes. Ordination is typically used to reduce variable dimensionality to a much smaller number of interpretable dimensions, to summarize variable inter-correlations, to determine the relative contribution of variables to the underlying data structure, and to quantify variable redundancy (Jeffers 1982). Interpretation of ordination axes typically results in the generation of hypotheses regarding possible causal factors determining the underlying data structure: for example, biotic data ordination axes are often interpretable in terms of underlying environmental factors or gradients. For this reason ordination is sometimes referred to as "indirect gradient analysis", under the assumption that the ordination axes correspond to underlying environmental factors indirectly measured by the sampled biota (ter Braak and Prentice 1988). However, it must be recognized that numerous additional factors determine biotic composition and abundance, such as biotic interactions, historical factors, dispersal limitations, and chance effects. The two major goals of multivariate analysis – the representation of species composition in a derived low-dimensional space, and gradient analysis – should therefore be viewed as distinct objectives that cannot be simultaneously achieved by a single ordination strategy (Dea'th 1999). Ordination is a method for the efficient representation of compositional trends, but cannot be expected to fulfill optimally the objective of gradient analysis.

Canonical methods are appropriate when the broad objective is gradient analysis. More generally, canonical methods are used to examine the relationship between two sets of variables measured on the same sampling units. Typically, the dependent or response variable set consists of species composition-abundance (biotic data set), while the second contains the independent or predictor variables, typically environmental factors (abiotic data set). Under this scenario, the objective is to determine the extent to which the environmental data determines or predicts biotic composition, and to obtain information as to which variables (both abiotic and biotic) contribute most strongly to the relationship. Such an approach is sometimes referred to as "direct gradient analysis", since the environmental factors thought to influence biotic composition are quantified and examined directly (ter Braak and Prentice 1988). However, to avoid confusion the term "direct gradient analysis" is used here only in its historical sense: the approach of fitting species response curves or surfaces to pre-specified environmental gradients (Whittaker 1978). Of course, canonical methods can also be used to examine the relationship between two biotic data sets (e.g., plant species and insect species), or two environmental data sets (e.g., soil variables and climatic variables).

DATA TRANSFORMATION

In both univariate and multivariate analysis, examination of the frequency distributions of variables is critical to the selection of an appropriate analytical strategy. While linear statistical methods implicitly assume that variables are normally distributed, the vast majority of biological and physical measurements are distributed log-normally; that is, frequency distributions are positively skewed to a greater or lesser extent (Limpert et al. 2001). Both normal and log-normal distributions arise as a consequence of a large number of factors acting independently on the variables. A normal frequency distribution occurs when factor interactions are additive, whereas multiplicative interactions result in a log-normal frequency distribution. The factors governing frequency distributions in nature are typically multiplicative – growth is proportional to present size, for example, indicating a relative rather than absolute scale – implying that biological and physical measurements should be analyzed on a logarithmic scale (Mead 1988). Fortunately, analysis on a

log-scale is easily achieved through logarithmic transformation of continuous variables prior to the application of linear statistical methods. Mead (1988) notes that logarithmic transformation of variables greatly aids in meeting the assumptions of linear models – including homogeneity (reduction in variability), normality (reduction in skewness), and additivity (conversion to a linear scale) – leading him to state: "the statistical moral is that it should be assumed that data for continuous variables should be transformed to a log scale". This sound advice applies equally well to multivariate data analysis (Digby and Kempton 1987).

The critical importance of data transformation generally, and the logarithmic transformation in particular, is often not fully appreciated by practitioners of multivariate data analysis. Continuous abiotic (environmental) variables must be log-transformed, both to increase normality and to ensure additivity (Mead 1988). The logarithmic transformation is particularly useful in reducing the influence of so-called "outliers" (which are actually natural and expected quantities, since multiplicative effects will occasionally produce very large values; see Limpert et al. 2001).

For multivariate biotic survey data, two distributional properties must be considered: the P frequency distributions of the individual species (abundance values of a given row in the data matrix), and a single distribution of total species-abundances (i.e., the distribution of the P row total values). The frequency distribution of enumerated abundance values for a single species over the N sampling units is typically strongly skewed to the right: there are a few large values, a few more intermediate values, a great many small values, and a greater or lesser number of zero values depending on whether the species is rare or common (Legendre and Legendre 1998). The main benefit of applying a log-transformation to such data is to reduce the effect of the large values, which would otherwise be perceived as "outliers" by linear models (Digby and Kempton 1987; Mead 1988).

Consider now the single species-abundance distribution, i.e., the distribution of species commonness and rarity. This is the distribution of row totals, i.e., computed total abundance for each species across the N sampling units. Numerous studies have demonstrated that species-abundance distributions are characterized by a few very abundant or common species, a few more at medium abundance, many more at low abundance, and a great many rare species at very low abundance (Fisher et al. 1943; Preston 1948; Whittaker 1965; Pielou 1975; Wilson 1991; Magurran 2004). Species-abundance data based on extensive surveys and very large sample sizes are distributed log-normally, but distributions based on a smaller sample size typical of most multivariate data sets are truncated [the "veil line" of Preston (1948)] and are generally indistinguishable from a log-linear (log-series, geometric or related) distribution (Tokeshi 1999; Hubbell 2001; Magurran 2004). In fact, multivariate biotic survey data are often characterized by the occurrence of a great many rare species (Digby and Kempton 1987; Legendre and Legendre 1998), which contribute substantially to the proportion of zero values in the data (e.g., a species occurring only once will contribute N-1 zero values to the data matrix). Some multivariate methods are particularly sensitive to, and the results unduly influenced by, the presence of rare species. This sensitivity has led to the development of ad hoc procedures for "down-weighting" the influence of rare species (ter Braak and Smilauer 2002), and to more extreme recommendations such as the wholesale elimination of rare species from the data prior to analysis (McCune and Grace 2002). Neither approach can be condoned on theoretical or biological grounds. Instead, practitioners should seek out ordination methodologies and strategies that implicitly recognize and account for the expected log-linear species-abundance distribution of biotic survey data. The logarithmic transformation is extremely useful in this regard, since it linearizes the species-abundance relationship (Tokeshi 1999; Margurran 2004). This in turn linearizes the weighted contribution of species in multivariate analysis. Without log-transformation, the analysis may be entirely dominated by a few abundant species (Digby and Kempton 1987; McGarigal et al. 2000).

A minor complication occurs in transforming variables containing zero values to a log scale, since the logarithm of zero is not defined. The simplest solution is to add an arbitrarily small number (equal to or less than the smallest observable value) to each data value prior to transformation (Mead 1988; Legendre and Legendre 1998; McCune and Grace 2002).

DATA STANDARDIZATION

Standardization of Variables

Environmental variables are usually measured on different scales, that is in different units. For example, soil data can include variables expressed as depths, concentrations, pH-units, and so forth. It is therefore necessary to standardize the variables prior to multivariate analysis to achieve commensurability. The most widely used method is standardization by the standard deviate, also known as conversion to variable "z-scores". This standardization renders the variables scale-free and dimensionless, with each having zero mean and unit variance (Legendre and Legendre 1998). This is the standardization "built into" the product-moment correlation coefficient. In multivariate analysis, the correlation standardization (i.e., conversion to z-scores) gives equal a priori weighting (variances) to the P variables.

With biotic survey data, all variables (e.g., species) are generally measured in the same units (e.g., counts). It is therefore not necessary to standardize the variables, and it is often undesirable to do so. Some researchers recommend standardizing species variables to z-scores to give equal weight to all species in multivariate analysis (e.g., McGarigal et al. 2000). While weighting species equally might seem reasonable and even desirable, the consequences of doing so must be considered in light of the log-linear species-abundance distribution characteristic of biotic survey data. Specifically, this standardization equalizes the contribution of common and rare species in the analysis, including very rare species (e.g., those occurring only once or twice) that provide very little "information" to the data. Such a standardization is therefore far too severe, producing results that are largely dominated by the chance occurrences

of rare species. Digby and Kempton (1987) note an additional disadvantage of standardization by the standard deviate. Consider two species with the same mean abundance in a sample data set. The first species occurs at similar – but randomly varying – abundance in all sampling units (low variance), while the second has consistently high cover in wetter areas and low cover in dry areas (high variance). The second species is therefore diagnostic of soil moisture conditions, while the first is not. A multivariate analysis based on unstandardized (or log-transformed) data would reveal the moisture gradient clearly, but standardization would equalize the variable variances and thus obscure the species-environment relationship.

In general, multivariate analysis results based on unstandardized variables are often dominated by abundant species, whereas those based standardized variables are dominated by rare ones. A compromise solution is to base the analysis on log-transformed data, which linearizes the species-abundance relationship (i.e., species weights). The logarithmic transformation is also a useful alternative to standardization by standard deviate, as it renders the data values scale-independent "save for an additive constant" (Digby and Kempton 1987).

Standardization of Sampling Units

It is sometimes necessary to standardize or relativize the sampling units, particularly when the time and/or effort available for data collection varies among them. For example, consider a set of light traps established to determine the composition and abundance of insect pollinators. The length of time over which insects are collected may vary among the traps. A common approach to this problem is to standardize the sampling units by their variable totals (i.e., summations over all species) to produce proportional or percentage data, also known as compositional data (Aitchison 1986). This standardization relativizes the data, constraining the sample space to a geometric simplex (i.e., the N sampling units are constrained to a space of P-1 dimensions). Standardization of sampling units by their vector lengths [e.g., chord distance, Legendre and Legendre (1998)] is similar, but produces a "curved" geometric simplex. A constrained simplex space is radically different from the standard Euclidean space of vector data, although this is not always appreciated. The analysis of simplex-constrained data requires alternative multivariate approaches such as log-ratio analysis and related techniques (Aitchison 1986; ter Braak and Smilauer 2002). Some data sets are naturally compositional: for example, soil texture is often quantified as percentages of sand, silt and clay (constrained to total 100%).

Double Standardization (Variables and Sampling Units)

The most commonly used double standardization is the contingency deviate, i.e., simultaneous standardization by row and column totals. When a double standardization is performed, it is implicitly assumed that the data can be viewed as a contingency table, i.e., that relative rather than absolute variation is of interest (Legendre and Legendre 1998). This standardization is "built into" the ordination technique known as correspondence analysis. The implications and assumptions of double standardization in the context of correspondence analysis are discussed in greater detail below.

Local Standardization

Some empirically derived distance coefficients have "built-in" local standardizations, meaning that the denominator changes with each paired comparison of sampling units. These coefficients are known as "semi-metrics" (Legendre and Legendre 1998) or "scrambled forms" (Orloci 1978), since local standardization necessarily distorts Euclidean vector space. These coefficients cannot be recommended for general use, based on both theoretical and statistical grounds (Digby and Kempton 1987). Some researchers have recommended the semi-metric "percent difference" (also known as the Bray-Curtis or Sorenson) coefficient for general use (e.g., Faith et al. 1987; McCune and Grace 2002). However, this coefficient is very sensitive to large outlying values, and it strongly distorts Euclidean vector space (Digby and Kempton 1987; Legendre and Legendre 1998). Local standardization is unnecessary and should be avoided.

ORDINATION METHODS

Ordination or scaling methods achieve an efficient and optimized low-dimensional representation of a complex data structure by emphasizing and bringing to the forefront underlying trended variation while suppressing "noise" (Gauch 1973). This objective is readily achievable since multiple variables normally show coordinated responses to one or more (typically many) underlying factors. Ordination methods are generally used in exploratory data analysis, both to search for and summarize underlying trends, and to examine interrelationships among variables. A number of ordination approaches are available, including principal component analysis (PCA) and its variants, correspondence analysis (CA), and non-metric multidimensional scaling (NMDS). The discussion here is meant to provide a basic understanding of the assumptions, advantages and limitations of available ordination methods and is therefore purposefully selective and non-technical. Statistical aspects of multivariate methods used in the biological sciences can be found in Digby and Kempton (1987), Legendre and Legendre (1998), McGarigal et al. (2000), Kenkel et al. (2002), and Gotelli and Ellison (2004).

Principal Component Analysis (PCA)

Principal component analysis (PCA) obtains an optimized, low-dimensional representation of coordinated variable response along mutually orthogonal ordination axes in Euclidean space. These ordination axes are derived through rigid (and therefore linear) rotation of the P variables such that the proportion of variance accounted for is maximized. The first axis maximizes linear variance (i.e., it summarizes the dominant linear trend), the second maximizes the residual variance not accounted for by the first axis (i.e., subdominant linear trend), and so forth. The method is somewhat analogous to simple linear regression, but extended to multiple dimensions and without distinguishing between dependent and independent variables (Kenkel et al.

2002). Principal component analysis and related methods do not actually reduce dimensionality, but instead optimally re-express linear variation along derived (variable-weighted) ordination axes. Dimension reduction is achieved by retaining only the dominant ordination axes, in much the same way that simple linear regression expresses the relationship between two variables as a "best-fit" line: variation about the linear trend is considered "error" and is generally ignored in predictive modeling. The first few ordination axes, which account for much of the total variance of the data set, will therefore normally provide a reasonable representation of the underlying linear trends of coordinated variable response (Gauch 1973).

Principal component analysis is a straightforward method of linear transformation, in which the total variance of the *P* variables is repartitioned along orthogonal ordination axes. The ordination axes are therefore derived variables expressed as linear combinations of the original *P* variables. Specifically, each variable is "weighted" on a given ordination axis in accordance to its contribution to the linear trend summarized by that axis. Consider, for example, a multivariate data set with a single linear underlying trend. The first ordination axis will summarize this trend, and species displaying a coordinated response to this trend will be highly weighted (either positively or negatively according to their relationship to the overall trend). By contrast, variables showing an uncoordinated response (i.e., trending away from the majority) will have low weightings.

Coordinate positions of the sampling units on the derived ordination axes are readily obtained as simple linear combinations of the variable weights: this is possible because ordination axes are obtained through rigid rotation of the original variable axes in Euclidean space. Ordination results are normally displayed as a scatter diagram (or scatterplot), showing the coordinate positions of the sampling units on the first two (i.e., most important) ordination axes. The display is rendered more interpretable by adding variable vectors (biplot scores), which are readily derived from the variable weights (Digby and Kempton 1987; Gower and Hand 1996).

Principal component analysis summarizes linear trends, and as such is the method of choice for the effective summarization of linear data structures. Continuous abiotic (environmental) survey data should always be analyzed using PCA (Leps and Smilauer 2003). Biotic survey data are also very often amenable to PCA, provided that the underlying data structure is broadly linear (that is, that the data do not encompass so great a range of environmental variation that species responses are non-linear). Many biotic survey data sets are broadly linear, making PCA the ordination method of choice in most studies. Principal component analysis also assumes variable normality; thus, like simple regression analysis it is sensitive to outliers. Log-transformation of continuous variables will eliminate the outlier problem (Digby and Kempton 1987).

Principal coordinate analysis (PCoA), also known as metric multidimensional scaling, is a generalized variant of PCA that has become increasingly popular in recent years (Podani and Miklos 2002). The method produces a mapping of sampling units, in which the pair-wise distances among sampling units in ordination space match as closely as possible their corresponding distances in variable space. A PCoA ordination based on Euclidean distances among sampling units is identical to that of a PCA based on a covariance matrix among variables, but the method can be generalized to accommodate other distance measures (Legendre and Legendre 1998; Podani and Miklos 2002). One drawback of PCoA is that variable weights (and thus biplot scores) are not produced, making it difficult to examine and interpret interrelationships between variables and ordination axes.

Another variant of principal component analysis, termed log-ratio analysis (Aitchison 1986), is appropriate when analyzing compositional (percentage) data with no, or few, zero elements (e.g., abiotic survey data). The most common approach to log-ratio analysis is loglinear-contrast principal components, which is the application of PCA to log-transformed data centered by both variables and sampling units (ter Braak and Smilauer 2002). Variable weights and biplot scores can also be determined in log-ratio analysis (Aitchison and Greenacre 2002). For compositional data containing many zeroes, correspondence analysis may be more appropriate (ter Braak and Smilauer 2002).

Correspondence Analysis (CA)

Correspondence analysis (CA) assumes a chi-square data space (versus the Euclidean data space assumed by PCA and its variants), and is therefore ideally suited to the analysis of contingency data (Greenacre and Hastie 1987). Reciprocal averaging (Hill 1973) and dual scaling (Nishisato 1980; Greenacre 1984) are theoretically identical to CA, although their statistical developments differ (Digby and Kempton 1987). As a method of dual (row-column) scaling, CA produces an ordination biplot in which the variables and sampling units are positioned according to their co-dependency (Jeffers 1982). Various CA algorithms scale the biplot scores differently, but this does not affect the interpretation of co-dependency (Legendre and Legendre 1998).

Curiously, CA can be viewed as either a linear or non-linear (unimodal) model (ter Braak and Smilauer 2002). Specifically, CA can be viewed as a special case of a linear PCoA in which the data are doubly standardized by the row and column totals (Digby and Kempton 1987). Alternatively, under specific limiting condition (see ter Braak and Prentice 1988), CA achieves a parsimonious representation of unimodal species responses to a single dominant environmental gradient.

PCA and CA are actually quite similar, both conceptually and statistically (Legendre and Legendre 1998). Whereas PCA operates in Euclidean data space to repartition the total variance as a series of optimized linear additive components (ordination axes), CA partitions the total contingency chi-square (or inertia) as a series of linear additive components within a chi-square data space (Digby and Kempton 1987). Derived PCA axes therefore maximize linear variation, whereas CA axes maximize the correspondence (or inertia) between the rows and column categories (variables and sampling units) of a data matrix. PCA summarizes trends by

finding lines of best fit (much like regression analysis), whereas CA summarizes trends by highlighting specific matrix co-occurrences (that is, co-dependency of row-column categories). In CA, a higher inertia value indicates a greater degree of correspondence between specific combinations of the variables and sampling units (Kenkel et al. 2002).

Because CA is a contingency-based multivariate approach, it is particularly sensitive to unique variable-sampling unit combinations. It is not uncommon for such co-occurrences to be highlighted at the expense of summarizing overall data trends (ter Braak and Smilauer 2002). This feature is most often a problem when CA is applied (or more correctly misapplied) to poorly structured biotic survey data, which are very often characterized by one (or a few) "aberrant" species values, i.e., species that occur at high abundance in one or a few sampling units but are otherwise uncommon or entirely absent in the remaining units. This situation rarely occurs when CA is applied to structured contingency data (e.g., Greenacre and Vrba 1984).

Non-Metric Multidimensional Scaling (NMDS)

Both principal component analysis (and its variants) and correspondence analysis utilize matrix algebra to derive unique, successive linear composites such that distances among sampling units in variable space are well represented in a low-dimensional ordination space. Non-metric multidimensional scaling (NMDS), or ordinal scaling, obtains a similar representation using only the rank order of distances, rather than the distance values themselves (Digby and Kempton 1987). NMDS thus maps the sampling units into an ordination space such that distances in the ordination space are ranked as similarly as possible to those in variable space. The theoretical advantage of this rank-order (ordinal) approach to ordination is that underlying assumptions of linearity (as in PCA and variants) or contingency (as in CA) are not required nor specified. The lack of underlying assumptions, however, necessitates the use of a computationally intensive iterative algorithm to derive an optimized ordination configuration. At each NMDS iteration the rank order relationship between ordination and variable space distances is improved through successive approximation. Iteration continues until the stress function, which measures the correspondence between ranked ordination and variable space distances, is minimized. The final solution is an optimized rank-order mapping of the sampling units in an ordination space of specified dimensionality. The solution obtained from a single run may not be globally optimal, however, since NMDS is based on an iterative algorithm. It is therefore imperative that multiple NMDS solutions be obtained in order to ensure that a stable and optimal ordination configuration is found. Because only rank order relationships are used, NMDS solutions are unstable or even degenerate when applied to small data sets or to poorly structured data.

NMDS has a number of additional disadvantages. Users must pre-specify the dimensionality of the ordination solution, which may be difficult given that the inherent underlying dimensionality of data is not generally known prior to analysis. An additional limitation relates to the comparison of ordination solutions in different dimensions. In PCA and CA, the r-dimensional solution is simply the first r dimensions of the s-dimensional solution, where $r < s$ (Digby and Kempton 1987). This is not true for NMDS, however, making it difficult to compare NMDS results across dimensions. Furthermore, NMDS ordination axes merely define a relative coordinate system and so cannot be interpreted in terms of their relative "importance" in summarizing the variation (as in PCA) or redundancy (as in CA) present in the data. Finally, a true ordination biplot cannot be produced in NMDS since variable weights are not determined, making interpretation much more difficult (Gotelli and Ellison 2004).

Given the many disadvantages and limitations of NMDS, it is not surprising that Digby and Kempton (1987) concluded that "we are unable to recommend the general adoption of non-metric methods to ecologists" (see also Legendre and Legendre 1998; McGarigal et al. 2000). Despite this, some researchers have recently championed the use of NMDS, based on its supposed robustness to departures from the "limiting" assumptions of PCA and CA (McCune and Grace 2002). While it is true that NMDS accepts a wide variety of distance measures (including non-linear and ordinal measures), in practice ordination solutions obtained using NMDS are rarely superior (and are very often inferior) to those obtained using PCA, CA and their variants (e.g., Digby and Kempton 1987; McGarigal et al. 2000). As with non-parametric univariate statistics (Sokal and Rohlf 1995), NMDS should be viewed as a method of last resort when applied to continuous or categorical data, to be used only when other analytical options have been exhausted. However, NMDS is more appropriate to the analysis of data consisting of variables measured on a rank-order (ordinal) scale (Podani 2005).

The Arch Effect

Simulation studies based on coenocline (single gradient) models have shown that the second ordination axis is a simple quadratic function of the first, producing two-dimensional ordinations in which the gradient is reproduced as an arch or distorted curve. This mathematical artifact, also known as the horseshoe or Guttman effect, is a natural consequence of applying most distance measures to single gradient data with unimodal species responses (Podani and Miklos 2002). The arch effect in both NMDS and PCoA can be alleviated using the method of flexible shortest path adjustment (Bradfield and Kenkel 1987; De'ath 1999). In CA, an empirically based method known as detrended correspondence analysis (DCA) is often used to remove the arch (ter Braak and Smilauer 2002). However, numerous studies have demonstrated that "detrending" may further distort ordination results (e.g., Jackson and Somers 1991). The general consensus in the recent literature is that DCA should be entirely avoided (Legendre and Legendre 1998; McCune and Grace 2002; Gotelli and Ellison 2004) or used with caution (McGarigal et al. 2000).

While problems related to the arch effect have received much analytical attention, they are rarely a concern in prac-

tice. The objective of ordination is to elucidate unknown (but anticipated) underlying gradients, not to recover a single obvious one. When a single environmental gradient predominates, the researcher is invariably aware of this and has purposefully sampled along the gradient to quantify species responses. The resultant data are therefore best analyzed using direct gradient analysis (sensu Whittaker 1978), in which species responses are plotted directly along one or more known gradients. With such data, ordination is neither necessary nor required (Digby and Kempton 1987). The presence of the arch effect in CA (or any other ordination method) indicates a single dominant underlying gradient, implying that the results should be presented as a single ordination axis (not a two-dimensional scattterplot); therefore, there is no need to "detrend" at all (Legendre and Legendre 1998).

CANONICAL METHODS

Canonical methods are used to determine the common structure, or correspondence, between two sets of variables (P_1 and P_2) measured on the same sampling units (Legendre and Legendre 1998). A typical application involves determining the degree and nature of the relationship between species and environmental variables in a common sample. Other examples include examination of the relationship between two species data sets (e.g., plants and insects), or between species groups and habitats. In all cases, the objective is to examine and summarize the common structure of the two data sets (Dray et al. 2003).

A number of canonical methods have been proposed, reflecting the many possible approaches for assessing the common structure of two data sets. Symmetric or descriptive canonical methods are correlation-based: neither data set takes on a response or predictive role. Asymmetric or predictive canonical models are regression-based approaches, in which one variable set (typically environmental data) takes on a predictive role while the other (species data) contains response variables. Canonical methods are also distinguished based on assumptions regarding the underlying structures of the two data sets. PCA-based canonical models assume that both data sets are linear, whereas in CA-based models one or both data sets are assumed to have unimodal (non-linear) responses. A final distinction is made between models based on multiple regression (classical canonical approach) and those based on partial least-squares regression (co-inertia approach).

Canonical Correlation Analysis (CCor)

Canonical correlation analysis (CCor) is a symmetric, descriptive method for examining the linear relationship between two data sets containing P_1 and P_2 variables, respectively, (Legendre and Legendre 1998). The method determines the common correlation structure of the two data sets by maximizing the squared correlation between pairs of derived linear canonical axes: one in species space, the other in environment space. A maximum of t pairs of canonical axes are found, where t is the lesser of P_1 and P_2. Like all canonical models, successive canonical axes are obtained subject to their being uncorrelated with those previously obtained (Kenkel et al. 2002).

Numerically, canonical correlation analysis involves two simultaneous multiple regressions (Dray et al. 2003). As a result, the number of variables P_1 and P_2 must be much smaller than the number of sampling units. This, together with the assumptions of linearity and analytical symmetry, limits the applicability of CCor in many survey studies. The number of variables can be reduced by first performing separate PCA ordinations on the two data sets, and then subjecting the scores from the major PCA axes to CCor (Kenkel et al. 2002). This data-reduction approach is numerically equivalent to inter-battery analysis (Tucker 1958), also known as PCA-PCA co-inertia analysis (Dray et al. 2003).

Redundancy Analysis

Redundancy analysis (RDA), also known as PCA with instrumental variables (Rao 1964), is the canonical or constrained form of PCA (Legendre and Legendre 1998). The method is asymmetric and predictive, since it maximizes predictions for a set of response variables (species data) given a set of predictive variables (environmental data). The method is essentially a PCA in which the sampling unit locations in species space are restricted to be linear combinations of the predictor or environmental variables (ter Braak and Smilauer 2002). The method, which is based on multiple linear regression analyses (one for each of the P_1 species on all the P_2 predictor variables), generally produces results similar to CCor. An example demonstrating that RDA is simply a principal component analysis of the fitted values obtained from the P_1 multiple linear regression analyses is provided by Legendre and Legendre (1998).

Redundancy analysis is the appropriate canonical model when both data sets are linear, and when an asymmetric analysis is required (i.e., when environmental variables are used to predict species composition, but not vice-versa). Because the method involves multiple regression, the number of predictor variables must be small relative to the number of sampling units and species: otherwise, the results are very similar to unconstrained PCA of the species data (Dray et al. 2003).

Canonical Correspondence Analysis

Canonical correspondence analysis (CCA) is the canonical or constrained version of CA, and is thus closely related to RDA (ter Braak and Prentice 1988; Legendre and Legendre 1998). Like RDA, the method uses multiple regression to obtain linear combinations of the predictor variables that best explain sampling unit positions in species space (ter Braak and Smilauer 2002). CCA is the appropriate predictive, asymmetric model when species responses are unimodal (response variables as contingency data) and the environmental (predictor) data are linear (ter Braak 1986). As in RDA, the number of predictor variables must be small relative to the number of sampling units.

CCA is undoubtedly the most widely used canonical model in ecology. This is largely attributable to the availability of proprietary CCA computer programs (Dray et al. 2003), and to the numerous studies advocating CCA for general use (Leps and Smilauer 2003). The method is a constrained form of CA and is therefore well-suited to the

canonical analysis of contingency data. However, CCA produces highly distorted results when applied to linear biotic data: RDA should be used instead (Dray et al. 2003).

Co-Inertia Analysis

Co-inertia analysis (COIA) refers to a generalized approach, based on partial least-squares regression, for analyzing the common structure of two data sets (Doledec and Chessel 1994; Dray et al. 2003). Co-inertia analysis is a symmetric, descriptive model that permits analysis of two data sets using various linear and contingency approaches. The approach includes many existing methods as special cases: for example, a PCA-PCA COIA (i.e., a model assuming that both the species and environmental data are linear) is inter-battery analysis (Tucker 1958), which is in turn closely related to RDA and CCor. Co-inertia analysis differs from the "classic" canonical models (CCor, RDA, CCA) in utilizing partial least-squares regression, rather than multiple regression, to summarize common structure. However, comparable canonical and co-inertia methods often (but not always) produce similar results: thus PCA-PCA COIA is related to RDA, and PCA-CA COIA is related to CCA (Dray et al. 2003). The symmetric form of co-correspondence analysis, a canonical method for comparing two contingency tables, is equivalent to a CA-CA COIA (ter Braak and Schaffers 2004).

Because COIA is based on partial least-squares regression, it places no restrictions on the number of variables that can be analyzed (unlike the classic canonical models). The co-inertia model is symmetric, and therefore descriptive rather than predictive. However, in practice equivalent symmetric and asymmetric models often give very similar results. Canonical models may offer superior performance when some environmental variables show low correlation, whereas co-inertia analysis provides superior results when environmental variables are highly correlated (Dray et al. 2003). The main advantage of co-inertia analysis is its flexibility – as a general model, it incorporates various analytical combinations for comparing two data sets. Furthermore, the model can be extended to the analysis of a series of paired tables: for example, to examine changes in species – environmental relationships over a temporal series (Thioulouse et al. 2004).

DISCUSSION

Historically, recommendations regarding the choice of appropriate ordination methodologies in the biological sciences were based on inductive reasoning. Specifically, the performance of various methods and standardizations were assessed using data generated from simulated coenoplane models (Whittaker and Gauch 1978). For some researchers, this remains the principal or sole approach to assessing multivariate methods and strategies (e.g., McCune and Grace 2002). When simulated data are used in this way, it is implicitly assumed that: (1) different methodologies can be objectively and completely assessed based solely on their ability to recover highly specific, idealized data structures; (2) theoretical considerations can be ignored when comparing methods and strategies; and (3) the criteria for comparison are sensible and objective for both species and environmental data.

There is certainly merit in using simulated coenoplane models to illustrate the efficacy and limitations of particular analytical methods. However, using such models as the principal or sole strategy for assessing multivariate methods can result in highly misleading and even erroneous conclusions, particularly when such models fail to reflect accurately the underlying structure of most environmental and biotic data sets (Minchin 1987b).

The most restrictive limitations of coenoplane models relate to their very conceptualization (Gauch and Whittaker 1972a, 1976). They are niche-based models that consider idealized, unimodal response curves of dominant species along one or two continuous, lengthy and orthogonal environmental gradients (Minchin 1987a, b). Such models are clearly not representative of the great diversity, or indeed the vast majority, of "real" environmental and biotic data sets. Biotic survey data rarely, if ever, adhere to such a restrictive model, and the model is entirely irrelevant for environmental data. Given such severe limitations, it is in hindsight remarkable that coenoplane models historically played such a predominant role in assessing the efficacy of ordination methods and strategies, and more remarkable still is that they continue to do so today. An unfortunate recent trend is the selective use of coenoplane models to dismiss summarily established ordination methods and to advocate non-metric multidimensional scaling as the "future of ordination", and as "one of the most defensible techniques during peer review" (McCune and Grace 2002). Such statements are indefensible given the unrealistic properties of the simulated coenoplane data on which they are based.

A further limitation of coenoplane models is that they make no provisions for the oft-observed log-linear frequency distributions of both species-abundance values (i.e., the row totals), and of values for individual species across the sampling units (i.e., the values in a given row). An important consequence of the log-linear distribution of species-abundance is that most species are rare and therefore contribute disproportionately to the number of zeros in biotic survey data. By contrast, the coenoplane model implicitly assumes that zero values indicate not rarity, but rather positions along an underlying environmental gradient that are beyond the tolerance limits of a species. Uncritical application of the coenoplane model leads to the conclusion that biotic survey data "becomes increasingly sparse [greater proportion of zeros] as the range of environment encompassed by the sample increases" (McCune and Grace 2002, p. 38). While this is strictly true, the converse is certainly not: a data set containing a high proportion of zeroes (typical of biotic survey data) does not necessarily indicate a wide range of environmental variation. Indeed, it is entirely possible that a high proportion of zeros simply reflects the expected distribution of commonness and rarity, even in a study site that is environmentally invariant. Despite this, it is often argued that a high proportion of zeros indicates a broad range of environmental variation (Legendre and Legendre 1998). More controversially, a high proportion of zeros is sometimes used in support of the erroneous argu-

ment that PCA and related linear models are inappropriate to the analysis of biotic survey data (McCune and Grace 2002).

In practice the proportion of zeros in a data set is a function of both the species-abundance relationship (i.e., the distribution of commonness and rarity) and species tolerance limits along environmental gradients, making it difficult to separate their relative effects. However, it is instructive to note that published biotic survey data sets (including those used in evaluating ordination methods, e.g., Digby and Kempton 1987; McGarigal et al. 2000; ter Braak and Smilauer 2002; McCune and Grace 2002) are almost invariably characterized by a few ubiquitous (or nearly ubiquitous) species, a few more moderately common species, and a great many rare species: exactly the distribution expected from a log-linear species-abundance relationship (Tokeshi 1999; Margurran 2004). The presence of one or more ubiquitous species in these data leads to the inescapable (and sobering) conclusion that the range of environmental variation in many biotic survey data sets is actually rather small – or at least not as great as often assumed – since at least some species occur over the entire environmental range sampled. This in turn implies that the large proportion of zeros characteristic of biotic survey data is a direct consequence of the log-linear species-abundance relationship, and has much less to do with species tolerance limits and unimodal response curves. The implication of this observation is of critical importance, for it indicates the need for a paradigm shift in the approach used to select an appropriate ordination method and strategy. Specifically, much greater attention needs to be devoted to the distributional properties of biotic survey data, and much less to unrealistic coenoplane models of species tolerance limits and unimodal response curves.

The criteria for selecting appropriate ordination methods and strategies should ultimately be based on the scientific method: observational and theoretical considerations of data structures, analytical objectives and methodologies are therefore of paramount importance (Orloci 1978; Legendre and Legendre 1998; Hubbell 2001). Such considerations should take precedence over empirical investigations based on inductive reasoning (i.e., general statements based on limited coenoplane simulation models) when evaluating ordination methods. Empirical investigations invariably produce erroneous recommendations regarding the utility of ordination methodologies: thus, statements such as "Ecological community data are, however, rarely amenable to PCA", or that "… there should be no regular application of CA to ecological community data" (both quotes from McCune and Grace 2002) are both misleading and wrong.

GENERAL RECOMMENDATIONS

The development of an appropriate multivariate analytical strategy for a given data set should proceed as a careful sequence of steps, in which results obtained at a given step determine subsequent ones (Jeffers 1982). Multivariate data analysis is thus a process of adaptive learning, in which decisions made at a given analytical stage direct subsequent steps and strategies. Before proceeding with a formal multivariate analysis, it is critically important to complete a detailed exploratory analysis of the data. Exploratory analysis is undertaken to elucidate and summarize distributional properties and underlying trends of the data, which in turn direct the user to meaningful analyses and interpretations (Tukey 1977; Legendre and Legendre 1998). Digby and Kempton (1987) provide an excellent example of how exploratory analyses of the Park Grass Experiment data provide critical insights regarding necessary data transformations and variable standardizations, and for selecting the most appropriate ordination and canonical methods.

It must be emphasized that there is no single "best" ordination or canonical method. Rather, the underlying data structure and study objectives must be considered when choosing an appropriate methodology. Selection of an appropriate multivariate analytical strategy must therefore be made on a case-by-case basis, and should never be based solely on simplified "recommendations" found in the literature. Nonetheless, some methods are clearly of limited use, while others have a broader appeal. For example, NMDS cannot be recommended for general use: it offers few if any advantages over PCA and CA, and has a number of serious disadvantages. Conversely, PCA – which has been much maligned in the past and recent ecological literature – should certainly be much more widely used.

Some strategies to aid in the selection of appropriate multivariate methods and analytical strategies are presented below. These are meant as general guidelines only, and it must again be emphasized that considerations of study objectives and data structure (obtained through exploratory data analysis) must ultimately drive decisions concerning data transformation, variable standardization, and selection of ordination and canonical methods. These guidelines are largely based on my experience over the past 20 years as a multivariate analysis consultant to graduate students and colleagues across a broad range of biological sub-disciplines, including agronomy and weed science, soil science, environmental science, plant and animal ecology, numerical taxonomy, oceanography, and biogeography. I have found that selection of an appropriate multivariate analytical strategy invariably transcends disciplines, and that the single most important factor is data structure: that is, whether one is analyzing continuous abiotic (environment), continuous biotic (species), or categorical contingency data.

Ordination Methods

As with univariate analysis, it is strongly recommended that multivariate analysis of categorical or continuous data proceed by first applying linear ordination methods having well-known underlying assumptions and statistical properties. Linear ordination methods (PCA and CA) are statistically powerful and remarkably robust to moderate deviations from underlying assumptions (Sokal and Rohlf 1995; Legendre and Legendre 1998). Non-metric ordination, like univariate non-parametric statistics, should only be used when all other options (including data transformation and variable standardization) are exhausted, and it can be convincingly demonstrated that the assumptions of more statistically robust models are clearly violated. Despite my

earlier advocacy of NMDS (Kenkel and Orloci 1986; Bradfield and Kenkel 1987), I have found that non-metric ordination is very rarely required.

Abiotic (Environmental) Survey Data

Analytically, this is certainly the most straightforward data structure. All continuous variables should be linearized through log-transformation except, of course, for those recorded on a logarithmic scale (e.g., soil pH). Since environmental data invariably consist of variables measured on different scales, variables must be standardized by standard deviation (*z*-scores). The appropriate method for this type of data is PCA of a correlation matrix based on log-transformed data. CA is inappropriate in this case because it will severely distort the underlying data structure by emphasizing relative rather than absolute differences (ter Braak, 1986; ter Braak and Smilauer 2002), and because the variables are measured on different scales (Leps and Smilauer 2003).

Biotic (Species) Survey Data

Species abundance values should be log-transformed prior to ordination, both to linearize the relative contributions of common and rare species and to alleviate problems associated with outlier values. In general, species variables should not be standardized to *z*-scores. In the rare case where species variables are measured on different scales, log-transformation may be a better choice than standardization to *z*-scores (Digby and Kempton 1987). An initial analysis of species survey data should employ PCA of a covariance matrix based on log-transformed data. This recommendation runs counter to that of many researchers advocating CA (or DCA) or NMDS for the ordination of biotic data, but in my experience most biotic survey data are broadly linear (i.e., they encompass a relatively modest range of environmental variation). In such cases interest focuses on absolute rather than relative differences in species abundance, implying that PCA is the correct model. Automatic application of CA or NMDS implies that biotic data are characterized by high species turnover (and unimodal response curves), but there is no prior reason to suppose this to be the case (Legendre and Legendre 1998; McGarigal et al. 2000).

The efficacy of a linear PCA model can be assessed through careful examination of a two-dimensional ordination biplot, which displays both the positions of sampling units and vectors of species correlations. The positions of sampling units relative to one another, and to the species biplot vectors, should be examined and related back to trends in the raw data (sorting the species and sampling units by their scores on the first ordination axis will simplify the comparison). Practitioners should ask: does the ordination configuration reflect well the underlying data structure? Ultimately, the "success" of the ordination model in summarizing underlying data structure must be determined by the individual researcher (Jeffers 1982; Legendre and Legendre 1998).

If the PCA ordination configuration reflects the underlying data structure, it is not necessary to seek an alterative model. Conversely, a poor ordination representation indicates that a non-linear model may be appropriate (contingency data, discussed below), or alternatively that the data are poorly structured. Poorly structured data arise when one or more variables are exclusive to one or more sampling units that contain only these variables – that is, disjunctions occur in the data set. Disjunct data, which are impossible to adequately represent by any ordination technique, often indicate inadequate sampling effort in surveying the biota. Examples of poorly structured data, and the analytical challenges they present, can be found in the recent monograph by Shaw (2003).

Contingency Data

Correspondence analysis (CA) is the ordination method of choice for analyzing contingency data. Since CA undertakes a simultaneous double standardization by row and column totals, variables should not be standardized to *z*-scores. A logarithmic transformation is generally not necessary and may be undesirable, since the double standardization of CA implies that relative rather than absolute differences are summarized.

Application of CA to contingency data may occasionally produce a two-dimensional ordination with a pronounced arch effect, indicating predominance of a single strong underlying gradient. If this occurs, the results should be summarized along the first ordination axis (Legendre and Legendre 1998). Alternatively, the environmental factor reflecting the gradient can be identified (it is usually obvious), and the species responses displayed using direct gradient analysis (Digby and Kempton 1987). It is rarely necessary, nor desirable, to utilize detrended correspondence analysis (DCA).

CA is highly sensitive to outliers, particularly unique row-column combinations (e.g., a sampling unit with high abundance of a species that is otherwise absent, or at low abundance, in all other sampling units). This problem most commonly arises when CA is applied to biotic survey data, which are often "sparse" (e.g., a given species may be dominant in only one sampling unit). For this reason, CA cannot be recommended for general use in the analysis of biotic survey data. However, CA is well-suited to the analysis of contingency data obtained from extensive biotic surveys that involve substantial sampling effort (e.g., Greenacre and Vrba 1984).

Canonical Methods

Canonical analysis should always be preceded by separate ordination analyses of the two data sets that are to be canonically compared (typically, a biotic or species data set, and an abiotic or environment data set). If this is done, the selection of an appropriate canonical method is very straightforward. Because environmental data are invariably linear (most often following logarithmic transformation), PCA is the appropriate model (ter Braak 1986). If the biotic data are also linear (i.e., the PCA model is deemed appropriate), then the canonical model of choice is RDA (or PCA-PCA COIA) using a covariance matrix of log-transformed species abundance data. For biotic contingency data, the appropriate canonical model is CCA (or CA-PCA COIA). Finally, co-

correspondence analysis (or CA-CA COIA) is the appropriate method for comparing two contingency tables.

ACKNOWLEDGEMENTS

This research was supported by the Natural Sciences and Engineering Research Council. I thank Rod Lastra, David Walker and two anonymous reviewers for their very helpful comments and criticisms.

Aitchison, J. 1986. The statistical analysis of compositional data. Chapman and Hall, London, UK.
Aitchison, J. and Greenacre. M. 2002. Biplots of compositional data. Appl. Statist. **51**: 375–392.
Bradfield, G. E. and Kenkel, N. C. 1987. Nonlinear ordination using flexible shortest path adjustment of ecological distances. Ecology **68**: 750–753.
De'ath, G. 1999. Principal curves: a new technique for indirect and direct gradient analysis. Ecology **80**: 2237–2253.
Digby, P. G. N. and Kempton. R. A. 1987. Multivariate analysis of ecological communities. Chapman and Hall, London, UK.
Doledec, S. and Chesssel, D. 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biol. **31**: 277–294.
Dray, S., Chessel, D. and Thioulouse, J. 2003. Co-inertia analysis and the linking of ecological data tables. Ecology **84**: 3078–3089.
Faith, D. P., Michin, P. R. and Blebin, L. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio **69**: 57–68.
Fisher, R. A., Corbet, A. S. and Williams, C. B. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol. **12**: 42–58.
Gauch, H. G. 1973. The relationship between sample similarity and ecological distance. Ecology **54**: 618–622.
Gauch, H. G. and Whittaker, R. H. 1972a. Coenocline simulation. Ecology **53**: 446–451.
Gauch, H. G. and Whittaker, R. H. 1972b. Comparison of ordination techniques. Ecology **53**: 868–875.
Gauch, H. G. and Whittaker, R. H. 1976. Simulation of community patterns. Vegetatio **33**: 13–16.
Gotelli, N. J. and Ellison, A. M. 2004. A primer of ecological statistics. Sinauer, Sunderland, MA.
Gower, J. C. and Hand, D. J. 1996. Biplots. Chapman and Hall, London, UK.
Greenacre, M. J. 1984. Theory and applications of correspondence analysis. Wiley, New York, NY.
Greenacre, M. J. and Hastie, T. 1987. The geometric interpretation of correspondence analysis. J. Am. Statist. Assoc. **82**: 437–447.
Greenacre, M. J. and Vrba, E. S. 1984. Graphical display and interpretation of antelope census data in African wildlife areas using correspondence analysis. Ecology **65**: 984–997.
Hill, M. O. 1973. Reciprocal averaging: an eigenvector method of ordination. J. Ecol. **61**: 237–249.
Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, NJ.
Jackson, D. A. and Somers, K. M. 1991. Putting things in order: the ups and downs of detrended correspondence analysis. Am. Nat. **137**: 704–712.
Jeffers, J. N. R. 1982. Modeling. Wiley, New York, NY.
Kenkel, N. C. and Orloci, L. 1986. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. Ecology **67**: 919–928.
Kenkel, N. C., Derksen, D. A., Thomas, A. G. and Watson, P. R. 2002. Multivariate analysis in weed science research. Weed Sci. **50**: 281–292.
Legendre, P. and Legendre. L. 1998. Numerical ecology. 2nd ed. Elsevier, Amsterdam, the Netherlands.
Leps, J. and Smilauer, P. 2003. Multivariate analysis of ecological data using CANOCO. Cambridge University Press, Cambridge, UK.
Limpert, E., Stahel, W. A. and Abbt, M. 2001. Log-normal distributions across the sciences: keys and clues. Bioscience **51**: 341–352.
Magurran, A. E. 2004. Measuring biological diversity. Blackwell, Oxford, UK.
McCune, B. and Grace, J. B. 2002. Analysis of ecological communities. MJM Software Design, Gleneden Beach, OR.
McGarigal, K., Cushman, S. and Stafford, S. 2000. Multivariate statistics for wildlife and ecology research. Springer, Berlin, Germany.
Mead, R. 1988. The design of experiments: statistical principles and practical applications. Cambridge University Press, Cambridge, UK.
Minchin, P. R. 1987a. An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio **69**: 87–107.
Minchin, P. R. 1987b. Simulation of multidimensional community patterns: towards a comprehensive model. Vegetatio **71**: 145–156.
Nishisato, S. 1980. Analysis of categorical data: dual scaling and its applications. University of Toronto Press, Toronto, ON.
Orloci, L. 1966. Geometric models in ecology. I. The theory and application of some ordination methods. J. Ecol. **54**: 193–215.
Orloci, L. 1978. Multivariate analysis in vegetation research. 2nd ed. Junk, the hague, the Netherlands.
Pielou, E. C. 1975. Ecological diversity. Wiley, New York, NY.
Podani, J. 2005. Multivariate exploratory analysis of ordinal data in ecology: pitfalls, problems and solutions. J. Veg. Sci. **16**: 497–510.
Podani, J. and Miklos, I. 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. Ecology **83**: 3331–3343.
Preston, F. W. 1948. The commonness and rarity of species. Ecology **29**: 254–283.
Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhya Ser. A **26**: 329–358.
Seal, H. L. 1964. Multivariate statistical analysis for biologists. Methuen, London, UK.
Shaw, P. J. A. 2003. Multivariate statistics for the environmental sciences. Arnold, New York, NY.
Sokal, R. R. and Rohlf, F. J. 1995. Biometry. 3rd ed. Freeman, New York, NY.
Swan, J. M. A. 1970. An examination of some ordination problems by use of simulated vegetational data. Ecology **51**: 89–102.
ter Braak, C. J. F. 1986. Canonical correspondence analysis a new eigenvector technique for multivariate direct gradient analysis. Ecology **67**:1167–1179.
ter Braak, C. J. F. and Prentice. I. C. 1988. A theory of gradient analysis. Adv. Ecol. Res. **18**: 271–317.
ter Braak, C. J. F. and Schaffers, A. P. 2004. Co-correspondence analysis: a new ordination method to relate two community compositions. Ecology **85**: 834–846.
ter Braak, C. J. F. and Smilauer, P. 2002. CANOCO reference manual and Canodraw for Windows user's guide. Microcomputer Power, Ithaca, NY.
Thioulouse, J., Simier, M. and Chessel, D. 2004. Simultaneous analysis of a sequence of paired ecological tables. Ecology **85**: 272–283.

Tokeshi, M. 1999. Species coexistence: ecological and evolutionary perspectives. Blackwell, Oxford, UK.
Tucker, L. R. 1958. An inter-battery method of factor analysis. Psychometrika **23**: 111–136.
Tukey, J. W. 1977. Exploratory data analysis Addison-Wesley, Reading, MA.
Whittaker, R. H. 1956. Vegetation of the Great Smoky Mountains. Ecol. Monogr. **26**: 1–80.
Whittaker, R. H. 1965. Dominance and diversity in land plant communities. Science **147**: 250–260.
Whittaker, R. H. 1978. Direct gradient analysis. Pages 7-50 *in* R. H. Whittaker, ed. Ordination of plant communities. Junk, The Hague, the Netherlands.
Whittaker, R. H. and Gauch, H. G. 1978. Evaluation of ordination techniques. Pages 277–336 *in* R. H. Whittaker, ed. Ordination of plant communities. Junk, The Hague, the Netherlands.
Wilson, J. B. 1991. Methods for fitting dominance/diversity curves. J. Veg. Sci. **2**: 35–46.