Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015, **6,** 399–411 doi: 10.1111/2041-210X.12236

SPECIAL FEATURE
NEW OPPORTUNITIES AT THE INTERFACE BETWEEN ECOLOGY AND STATISTICS

Model-based approaches to unconstrained ordination

Francis K.C. Hui [1,2,*], Sara Taskinen[3], Shirley Pledger[4], Scott D. Foster[2] and David I. Warton[1,5]

[1]*School of Mathematics and Statistics, The University of New South Wales, Sydney, NSW, Australia;* [2]*CSIRO Computational Informatics, Australia and CSIRO's Wealth from Oceans Flagship, Hobart, Tasmania, Australia;* [3]*Department of Mathematics and Statistics, University of Jyväskylä, Jyvaskyla, Finland;* [4]*School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, Wellington, New Zealand; and* [5]*Evolution & Ecology Research Centre, The University of New South Wales, Sydney, NSW, Australia*

Summary

1. Unconstrained ordination is commonly used in ecology to visualize multivariate data, in particular, to visualize the main trends between different sites in terms of their species composition or relative abundance.

2. Methods of unconstrained ordination currently used, such as non-metric multidimensional scaling, are algorithm-based techniques developed and implemented without directly accommodating the statistical properties of the data at hand. Failure to account for these key data properties can lead to misleading results.

3. A model-based approach to unconstrained ordination can address this issue, and in this study, two types of models for ordination are proposed based on finite mixture models and latent variable models. Each method is capable of handling different data types and different forms of species response to latent gradients. Further strengths of the models are demonstrated via example and simulation.

4. Advantages of model-based approaches to ordination include the following: residual analysis tools for checking assumptions to ensure the fitted model is appropriate for the data; model selection tools to choose the most appropriate model for ordination; methods for formal statistical inference to draw conclusions from the ordination; and improved efficiency, that is model-based ordination better recovers true relationships between sites, when used appropriately.

Key-words: correspondence analysis, latent variable model, mixture model, multivariate analysis, non-metric multidimensional scaling

Introduction

A difficulty in analysing multivariate data in ecology is visualizing the data in a concise, low-dimensional form, a problem for which ordination is commonly used, for example Quinn & Keough (2002 *et al.*, Chapters 17–18), Legendre & Legendre (2012, Chapter 9). A prime example is visualizing the main trends between different sites in terms of their species composition. Ordination aims to reduce data from many response variables (often hundreds of species) to just two (usually), so that sites can be plotted on a standard scatterplot to look for patterns between sites. In this study, we consider ordinations using the response data alone, without input from predictor (or 'environmental') variables. This is often referred to as *unconstrained ordination* (Clarke 1993) or *indirect gradient analysis* (ter Braak & Prentice 1988). We restrict attention to multivariate abundance or presence–absence data sets, and visualizing sites such that distances between them correspond as well as possible to their relative positions along some underlying ecological gradient.

By far, the most common method of unconstrained ordination in ecology is non-metric multidimensional scaling (nMDS, Kruskal , 1964a, b), an algorithm-based technique which iteratively repositions points until the relative distances in the ordination have the strongest possible monotonic fit to the pairwise dissimilarities between sites, as computed using some dissimilarity measure of the (possibly transformed) data. A related method, known as principal coordinate analysis (PCoA, Gower 1966), constructs an ordination by applying a singular value decomposition to a pre-prepared matrix of *a priori* pairwise dissimilarities. PCoA forms the basis for a number of multivariate methods currently used in ecology (Legendre & Anderson 1999; Anderson 2001; Anderson & Willis 2003), all of which can be understood as fitting linear models to principal coordinates. One particular choice of PCoA is correspondence analysis (CA, Hill 1974), which can be understood as PCoA using the chi-squared dissimilarity measure along with some pre-specified scaling of site and/or species totals (see also ter Braak 1985, for a model-based view of CA). Detrended corre-

*Correspondence author. E-mail: fhui28@gmail.com

spondence analysis (DCA, Hill & Gauch 1980) is a modification of CA to remove the 'arch effect' often seen in resulting ordinations (but see Legendre & Legendre 2012, pp. 471–472) . All of these methods are algorithm-based techniques to ordination, in that they process the data in a series of steps (an algorithm) to extract meaning without reference to some underlying statistical model.

The question of which dissimilarity measure to use and how to transform data prior to its use is critical to the results of ordination, and has been subject to much discussion in literature (e.g. Faith, Minchin & Belbin 1987; Anderson *et al.* 2011). However, it remains the case that guidance to ecologists on this core issue is in general terms only, with little by way of formal diagnostic checks to ensure these choices are adequate. Thus, the decisions of what dissimilarity measure and data transformation to apply are made on Faith, Minchin & Belbin (1987), that is, past empirical performance instead of using the data itself to guide and validate these decisions.

One potential consequence of choosing an inappropriate transformation and/or dissimilarity measure is the failure to appropriately account for the mean–variance relationship, a key property of multivariate abundance data. The mean–variance relationship is an important driver of the statistical properties and thus the performance of multivariate analysis including ordination. Transforming data can help to mitigate the effect of the mean–variance relationship, but it typically cannot remove it completely when there are many zeros, as is characteristic of abundance data in ecology. If the mean–variance relationship is not properly accounted for, trends in location are confounded with changes in dispersion, leading to potentially misleading results (Warton, Wright & Wang 2012).

The most direct way to ensure an appropriate mean–variance relationship is to account for it for each species. Generalized linear models (GLMs, McCullagh & Nelder 1989) are routinely used to model data with a mean–variance relationship. Moreover, various extensions of GLMs have been applied to regression of multivariate data in ecology, for example vector generalized additive models (Yee 2010), species archetype models (Dunstan *et al.* 2013), generalized estimating equations (Wang *et al.* 2012) and their effectiveness has been verified by various means (Warton, Wright & Wang 2012; Hui *et al.* 2013). Such models are a convenient means of separating out the signal from the noise – the model has a systematic component (linear predictor) and a stochastic component (distributional assumption). The two could not so readily be teased apart without such an explicit model for stochastic variation.

While model-based approaches to regression of multivariate abundance data have been developing rapidly, relatively little progress has been made in model-based approaches to ordination. Historically, some attention was given to modelling how species' mean response varies as a function of environmental variables, giving rise to Gaussian ordination (Gauch, Chase & Whittaker 1974; Ihm & van Groenewoud 1975). However, fitting these models proved challenging and computational difficulties were often encountered (Kooijman 1977), so correspondence analysis was often preferred, which can actually be understood as a first-order approximation to Gaussian ordination for both short gradients (Ihm & van Groenewoud 1984) and long gradients with equal tolerances among species (ter Braak 1985). Perhaps one reason for the computational difficulties was problems in model specification – ordination scores were typically treated as fixed in Gaussian ordination, when they would more naturally be treated as unobserved random or 'latent' variables.

Model-based approaches to constrained ordination are free of this issue, and software has been developed using maximum likelihood estimation assuming data arise from one of several common types of distribution (`VGAM` package in `R`, Yee 2010). No corresponding procedures have been developed for the problem of unconstrained ordination, beyond the special case of presence–absence data (Walker & Jackson 2011).

In this study, we propose two model-based approaches to unconstrained ordination, each capable of handling different data types and different forms of species response to latent gradients. Both can be regarded as extensions of GLMs and thus explicitly model the mean–variance relationship in the data. Furthermore, as model-based approaches, we can make use of many standard tools (developed for regression analysis) when performing an ordination. These include the following: residual analysis tools to check the assumptions made are suitable for the data; model selection tools for choosing the most appropriate model for ordination; and inference methods which allow one to draw generalizable conclusions from the ordination. We present examples and simulations to illustrate our model-based approaches to ordination, and compare the results obtained to those from algorithm-based techniques.

New approaches to unconstrained ordination

We propose two model-based approaches to unconstrained ordination – the first is an extension of mixture modelling approach proposed recently by Pledger & Arnold (2014), and the second uses a latent variable model (Skrondal & Rabe-Hesketh 2004). The two models differ in structure, which can provide insights into different aspects of the data set.

Throughout the following developments, it is assumed the multivariate abundances have been collected on p species and are correlated across species. It is this correlation we wish to exploit when constructing an ordination. The data are assumed to come from n independent sites, for example cross-sectional data at a given time from randomly sampled locations (although the below models could be extended to handle spatial or temporal autocorrelation). These data are typically arranged in a $n \times p$ matrix, with sites in rows and species in columns).

ORDINATION VIA MIXTURE MODELLING

Finite Mixture Modelling (McLachlan & Peel 2004) is a model-based method for unsupervised classification, which can be used to classify sites or species. A classification of species can be understood as a form of dimension reduction, and Pled-

ger & Arnold (2014) used this dimension reduction as motivation for a novel approach to unconstrained ordination.

Pledger & Arnold (2014) classify each of p species into one of a small number, say C, of clusters, using the following mean model:

$$g(\mu_{ij|c}) = \alpha_i + \beta_j + \gamma_{ic}, \quad i = 1,\ldots,n;\ j = 1,\ldots,p;\ c = 1,\ldots,C, \qquad \text{eqn 1}$$

where $\mu_{ij|c}$ is the mean response (count or probability of presence) for species j at site i, conditional on the species belonging to cluster c. Conditional on group membership, eqn (1) is a GLM, where a link function $g(\cdot)$ is used to relate the mean response to a set of covariates. For presence–absence data, we used the logit link function. For count data, we used the log link function.

This is a mixture model which mixes on species according to their site profile. Specifically, we assume that there are C types of column patterns or 'site profiles' specifying how relative abundance changes across sites. Each column (species) is then assumed to be drawn from one of these C types of site profiles. This model makes no assumptions with respect to response to any underlying gradients across sites – thus, it can be understood to be capable of fitting unimodal responses (as described in ter Braak & Prentice 1988) or indeed any other type of response to underlying gradients.

A critical aspect of the model is the statistical distribution assumed for the data and the mean–variance assumption this implies. Conditional on the species belonging to cluster c, we assumed a Bernoulli distribution (with mean–variance $Var(y_{ij}) = \mu_{ij|c}(1-\mu_{ij|c})$) for presence–absence data. For count data, we assumed either a Poisson ($Var(y_{ij}|c) = \mu_{ij|c}$) or negative binomial distribution ($Var(y_{ij}) = \mu_{ij|c} + \phi_j\mu^2_{ij|c}$ where ϕ_j are species-specific overdispersion parameters). While the Bernoulli assumption is, by definition, satisfied for presence–absence data, distributional assumptions for count data should be checked using diagnostic tools to ensure that the mean–variance relationship has been appropriately satisfied. We illustrate such checks in our worked example in Section 'Spider data set'

In eqn (1), the term β_j accounts for the differences in mean (and hence variance) across species, due to their different abundance (or for presence/absence data, their different species relative frequency). This is what standardizing variables in nMDS attempts to do. Two different ordinations of the sites arise from the model in eqn (1). In the first option, the vector $\gamma_i = (\gamma_{i1},\ldots,\gamma_{iC})$ is constrained to sum to zero, that is $\gamma_{i1}+\gamma_{i2}+\ldots+\gamma_{iC} = 0$ for all $i = 1,\ldots,n$. This implies the term α_i represents site effects (rich versus poor), while the vectors $\gamma_i = (\gamma_{i1},\gamma_{i1},\ldots,\gamma_{iC})$ can be interpreted as deviations from the main effects of site and species, that is, changes in species composition over the sites, after allowing for differences in site total abundance and species prevalence. If $C = 3$ clusters of species are used, then γ_i provides ordination coordinates for a *two*-dimensional scatter plot of the sites. This reduction in dimension arises from the sum-to-zero constraints; the points are in the plane $x+y+z = 0$ which is rotated into the x-y plane for a plot in two dimensions.

The second option from eqn (1) is to absorb the α_i term into the γ_{ic} elements,

$$g(\mu_{ij|c}) = \beta_j + \gamma_{ic}, \quad i = 1,\ldots,n;\ j = 1,\ldots,p;\ c = 1,\ldots,C,$$

with no constraints on γ_i. If $C = 2$ clusters of species are used, then the vector γ_i provides coordinates for a two-dimensional ordination of sites that are in terms of relative species abundance (an aggregate of site total abundance and species composition).

Whether or not the site effects should be included in the model separately or subsumed into the ordinations depends to some extent on the ecological question of interest. In particular, it depends on whether compositional change alone is of interest or changes in total abundance (species richness) also. In cases where either is applicable, we can let the data itself decide on the best approach using model selection tools. We can similarly treat choice of the number of clusters C as a model selection problem. This is discussed further in Section 'Advantages of model-based approaches'.

The mixture model in eqn (1) is fitted using maximum likelihood methods via the Expectation–Maximization algorithm (EM, Dempster, Laird & Rubin 1977). Mathematical details regarding the fitting procedure may be found in Pledger & Arnold (2014). The R code for fitting these models is freely available at http://homepages.ecs.vuw.ac.nz/~shirley/ under 'Mixture Model Ordination'.

ORDINATION VIA LATENT VARIABLE MODELS

The mixture model (MM) approach of Pledger & Arnold (2014) produces an ordination indirectly by classifying species. Here, we propose an alternative to model-based ordination which directly captures the position of sites along some underlying ecological gradient. This approach is based on latent variable models (LVMs). One type of LVM that may be familiar to readers is factor analysis (Knott & Bartholomew 1999), although a key difference between the model we propose below and a two-dimensional factor analytic model is that we do not assume multivariate normality. Instead, we assume a known mean–variance relationship as was the case with MMs. That is, we assumed a Bernoulli distribution for presence–absence data, and either a Poisson or a negative binomial distribution for count data. Walker & Jackson (2011) recently proposed this approach for ordination of presence–absence data in ecology, under the name 'random effects ordination', the main differences here being: our model and code are applicable to general GLM families; our model optionally includes site effects; and Walker & Jackson (2011) extended their model to handle quadratic responses, which is a possible next step for our model.

In LVMs, the mean response for species j at site i can be explained by a set of q underlying or 'latent' variables, as specified by the following mean model,

$$g(\mu_{ij}|\mathbf{z}_i) = \alpha_i + \beta_j + \mathbf{z}_i'\boldsymbol{\theta}_j, \quad i = 1,\ldots,n;\ j = 1,\ldots,p, \qquad \text{eqn 2}$$

where $z_i = (z_{i1},\ldots,z_{iq})$ is the q-dimensional vector of latent variable values at site i, each of which is assumed to be independent and standard normal in distribution. For the purposes of ordination, we typically use $q = 1$ or 2 latent variables in eqn (2). Conditional on the (unknown) latent variables, the model is a GLM, assuming that species are linearly related to the underlying latent variables. Note however that this model structure does offer some capacity to handle nonlinearity, in part from the site effects α_i, as discussed later. The species-specific intercepts β_j are analogous to those in MMs in eqn (1), that is, they adjust for variability in species due to prevalence.

An ordination is produced from the latent variables $\mathbf{z}_i$. If $q = 2$, the latent variables for each site are a pair of coordinates representing the position of the site in a two-dimensional ordination. Also, the vector $\boldsymbol{\theta}_j = (\theta_{j1},\ldots,\theta_{jq})$ are coefficients which quantify how each species' response is related to the latent variables, and could potentially be added to the ordination (of $\mathbf{z}_i$) to construct a biplot, giving an indication of how species composition differs across sites.

While the model as written above is linear, it has some capacity to handle nonlinearity and unimodality in response of species to latent variables, for two reasons. Firstly, nonlinearity can be handled within the confines of any linear model by adding polynomial terms (e.g. x, x^2) to the linear predictor, and the same process applies here – additional latent variables in the model could characterize nonlinearities in response to some underlying gradient. If present, this effect could be diagnosed as an 'arch effect' in ordinations. Secondly, and as pointed out by Cajo ter Braak in review, the site effects α_i offer some capacity to handle nonlinearity, as long as it takes a common form across all species. For example, if all species had a quadratic, unimodal response to the gradient with equal tolerance, the α_i could account for this by taking values that are a quadratic function of position along the underlying gradient.

In our LVM, we treated α_i as fixed effects, but they could instead be chosen to be random (e.g. as in Jamil & ter Braak 2013). Treating the site effects as random has the advantage of reducing the number of model parameters, so that it would not increase as sample size increases. This has some theoretical and computational benefits, but also introduces new challenges. For now, we leave the consideration of random site effects to future work.

In the LVM formulation above, the inclusion of both site α_i and species β_j effects ensures the ordination coordinates $\mathbf{z}_i$ are standardized for site total abundance and species prevalence, respectively, that is, the ordination is in terms of species composition. But as for MMs, eqn (2) can be modified to exclude site effects,

$$g(\mu_{ij}|\mathbf{z}_i) = \beta_j + \mathbf{z}_i'\boldsymbol{\theta}_j, \quad i = 1,\ldots,n;\ j = 1,\ldots,p,$$

in which case, the ordination of sites is in terms of their species abundance (an aggregate of site total abundance and species composition). As previously, the choice of whether to plot an ordination of abundance or composition can be viewed as an ecological one, but if desired, we can actually use a data-driven approach to decide whether or not to include site effects in a LVM, using model selection tools as in Section 'Spider data set'. One point of difference in the LVM case, however, is that dropping the α_i restricts the capacity of the model to absorb nonlinearities in response to a latent gradient. This issue does not apply for MM, which fits a model that already handles nonlinearity.

In practice, LVMs tend to be robust to violations of the normality assumption for the latent variables (Seong 1990; Wedel & Kamakura 2001). Indeed, the method is capable of producing ordinations which are quite non-normal in appearance, suggesting the normality assumption does not impose any strong constraint on analyses, as will be seen later in Section 'Worked examples'. This robustness to violation of normality is analogous to the situation encountered in Bayesian analyses, where the form of the posterior can be different to that of the prior. The additional constraints that the latent variables have mean zero and unit variance do not impose any restrictions on the model, as the species coefficients ($\boldsymbol{\theta}_j$) control the scaling. As is the case for nMDS, the location, scale and rotation of a latent variable ordination are arbitrary, and the purpose of the zero mean and unit variance is to fix the location and scale of the ordination, respectively (see Skrondal & Rabe-Hesketh 2004, Chapter 5). An additional constraint needs to be placed on the $\boldsymbol{\theta}_j$'s to fix the rotation of the ordination (for details see Appendix S1).

We fitted LVMs via maximum likelihood estimation, using the EM algorithm (also) in conjunction with Monte Carlo Integration. Details on the fitting procedure as well as example R code are provided in Appendices S1 and S6, respectively. Once fitted, we use the posterior modes as predicted values of $\mathbf{z}_i$ as the ordination coordinates, as is standard in latent variable modelling (see Skrondal & Rabe-Hesketh 2004, Chapter 7).

A COMPARISON OF THE MODEL-BASED METHODS

Two model-based approaches to unconstrained ordination have been presented above. As both are situated within a modelling framework, LVMs and MMs share some common strengths, such as formal approaches to model selection and model checking using residual analysis. The approaches however differ in two key respects: assumptions and computation.

Assumptions

A mixture model imposes assumptions on species – assuming there are a small number of 'species clusters', into which species are classified in a soft or probabilistic manner. This model is consistent with the observation that many species coexist in seemingly similar environmental conditions (for some possible mechanisms, see Scheffer & van Nes 2006). In contrast, latent

variable models make assumptions on sites – assuming there is a continuous latent gradient along which each site lies. Sites are assumed to be normally distributed along the gradient, and species are assumed to be linearly related to position along this gradient, although as previously mentioned the method is robust to violation of these assumptions. MMs can also be understood as a type of latent variable model where the latent variables are multinomial in distribution. Thus, the key difference between the two models, from a statistical perspective, can be reduced to whether we assume a multinomially distributed latent variable across species (MMs), or a normally distributed latent variable across sites (LVMs). If one or other of these sets of assumptions was not well supported by the data, this would typically become apparent during analysis, for example in model selection, or by the data suggesting a large number of clusters C or latent variables q were required for a reasonable model fit.

Computation

LVMs take a longer time to fit compared to MMs, because the likelihood function we wish to maximize does not have a closed form expression and needs to be estimated by numerical integration (see Appendix S1 for details). The difference in computation time can be substantial although it is not an issue for small data sets – for instance fitting a MM to the coral data set (see the worked example in Section 'Coral dataset') with two clusters took less than a one minute, whereas fitting the LVM with two latent variables and site effects took 1–2 min.

With the differences between LVMs and MMs, the question arises as to which method to choose. Fortunately, a key benefit of model-based approaches to ordination is the availability of model selection tools to help users select the best model for the data of interest. This advantage as well as others is discussed in the next section.

Advantages of model-based approaches

There are a number of advantages to specifying a statistical model for your data and estimating it via formal probabilistic methods such as maximum likelihood. We now highlight some of these benefits below.

CONTROLLING SPURIOUS DATA PROPERTIES

Model-based approaches allow the analyst to explicitly account for key statistical properties of the data. A prime example of this is the strong mean–variance relationship typically seen with count data. When not properly controlled for, trends in location (mean abundance) may be confounded with changes in dispersion (variance), leading to potentially misleading results in nMDS ordinations (Warton, Wright & Wang 2012 and Fig. 3a). By explicitly modelling the mean–variance relationship and checking that it has been adequately captured, LVM and MM ordinations are better able to uncover the true underlying pattern, when their underlying assumptions are reasonable.

A second example is when primary interest is in community composition and effects on site absolute abundance are considered spurious. Typically to control for abundance, practitioners either apply a row-standardization, that is, dividing all abundances at a site by their row sum, or use a dissimilarity measure that implicitly applies some row-standardization such as the Bray–Curtis distance. These methods are problematic though, because the sampling distribution of proportions (which characterize composition) changes as total abundance changes. With model-based approaches to ordination, differences in site abundance can be easily controlled for by including a site effect, that is α_i in eqns (1) and (2). This idea is illustrated in our worked example on the coral data set in Section 'Coral dataset'

MODEL CHECKING

As with other regression techniques such as GLMs, we can use residual analysis to check whether LVM or MM adequately fits the data at hand. Note that for presence–absence and count data, residual analysis is somewhat challenging due to the discreteness of the data. However, this difficulty can be overcome through the use of the residuals proposed by Dunn & Smyth (1996), which we refer to as Dunn–Smyth Residuals, defined for the (i,j)th observation (Y_{ij}) as:

$$r_{ij} = \Phi^{-1}\Big(u_{ij}F_{ij}(y_{ij}) + (1-u_{ij})F_{ij}^{-}(y_{ij})\Big),$$

where $\Phi(\cdot)$ and $F_{ij}(\cdot)$ are the cumulative probability functions of the standard normal distribution and Y_{ij}, $F_{ij}^{-}(y)$ is the limit as $F_{ij}(y)$ is approached from the negative side, and u_{ij} has been generated at random from the standard uniform distribution. Each u_{ij} introduces 'jittering' to discrete variables, spreading the probability mass evenly across the interval between $F_{ij}(y)$ and its previous value.

The main goal of residual analysis is to check our model assumptions. For instance, plots of residual against predicted response can be used to check for overdispersion and potential outliers. Quantile plots of the residuals can be used to assess goodness-of-fit.

MODEL SELECTION AND INFERENCE

In conjunction with model checking, we can use a range of well-developed model selection tools (developed for regression analyses) to select key aspects of our ordination. These include (but are not limited to) hypothesis testing, cross-validation, and information criteria such as AIC (Akaike 1974) and BIC (Schwarz 1978). In our worked examples, we use information criteria along with model checking to select how many clusters (for MMs) or latent variables (for LVMs), whether to use the Poisson or negative binomial distribution, and so on. Of course, the precise choice of model used for ordination is informed not just by data properties, but also by the analysis goal. For example, whether or not we include a site effect depends on whether we want to perform an ordination of sites

according to species composition only or species abundance (an aggregate of site total abundance and composition).

Inference with model-based approaches is also relatively straightforward. For instance, if we are interested in understanding how species' response changes along the latent 'species composition' surface, then we can examine the point estimates for the coefficients θ_j in a fitted LVM, as well as calculate confidence intervals for these estimates to provide an assessment of uncertainty, although the accuracy of such confidence intervals in this context is in need of evaluation.

EFFICIENCY

Using a model-based approach to ordination, one can expect to obtain more accurate estimates of the underlying site patterns if the proposed ordination model has similar attributes to the process that generated the data. This is a theoretical property in statistics known as efficiency. In particular, maximum likelihood estimation, which we use in this study to fit LVMs and MMs, is known to have attractive efficiency properties in large samples (e.g. Van der Vaart 2000). While such theoretical results are derived asymptotically in large samples, a common experience is that they provide good performance in small samples also. The efficiency of model-based ordination approaches for small samples is illustrated via simulation in Section 'Simulation studies'.

Worked examples

We provide two worked examples illustrating how to apply the model-based ordination methods proposed. Both data sets are available from the `mvabund` package (Wang *et al.* 2012). As well as directly showing the general process used to fitting and performing inference on these models, the examples serve to highlight some advantages of model-based approaches to ordination discussed in Section 'Advantages of model-based approaches'. In both examples, we used BIC to conduct model selection, BIC = $-2\times$log-Likelihood+ log $(n)\times$(#of parameters). Formulas for calculating the number of parameters in MMs and LVMs are provided in Appendix S2.

SPIDER DATA SET

The first data set contains counts of $p = 12$ species of spiders recorded at $n = 28$ sites (Van der Aart & Smeenk-Enserink 1974). The data set was used as an example in ter Braak (1986). We fitted MMs and LVMs with the number of clusters/latent variables ranging from zero to five, with Poisson or negative binomial distributions for the counts, and models with and without site effects. The case of zero clusters or latent variables is a special one – it implies that species are independent with mean response dependent only on species and/or site main effect. Thus, by including it in the model selection procedure, we are effectively testing whether there is an association between species across site.

Using BIC, the best MM used the negative binomial distribution and two clusters, while the best LVM used the negative binomial distribution and two latent variables with site effects (Table 1). The inclusion of site effects in the best LVM suggests different degrees of site total abundance, which should be adjusted for prior to looking for patterns in species composition. As discussed previously, one reason this could arise is nonlinearity in environmental response, although further analyses lend little support to this explanation (Warton 2008, Fig. 4). Also, the best LVM had a smaller value of BIC compared to the MM, suggesting the gradient model with site effects implied by the LVM was more appropriate for this particular data set compared to the species classification assumed by the MMs without site effects.

Table 1. BIC values for MMs and LVMs fitted to the spider data set. Model selection considered whether the distributional assumption for the counts was negative binomial (NB) or Poisson, whether or not to include site effects, and how many clusters or latent variables were required. The best models, in terms of lowest BIC, are shown in bold

	Number of Clusters/Latent Variables					
	0	1	2	3	4	5
MM – Poisson	7164	3479	2442	2234	1907	1933
MM – NB	2009	1783	**1593**	1610	1620	1604
LVM – Poisson						
No site	7164	2940	1824	1719	1709	1925
Site	4728	3099	1807	1734	1864	1910
LVM – NB						
No site	1783	1601	1593	1622	1625	1691
Site	2009	1602	**1578**	1616	1652	1755

For both MMs and LVMs, assuming a negative binomial distribution for the counts led to substantially lower values of BIC compared to Poisson distribution (Table 1), suggesting there was strong overdispersion in the spider data. This was further evidenced when we analysed the residuals. Plots of Dunn–Smyth residuals against linear predictors for the best LVM showed no obvious pattern (Fig. 1a), whereas the corresponding Poisson LVM with site effects and two latent variables exhibited a fan-shaped pattern indicative of overdispersion (Fig. 1c). Normal quantile plots of the residuals also indicated that the negative binomial LVM better satisfied the normality assumption of the residuals compared to the Poisson LVM (not shown). Residual plots for the best MM showed similar patterns (see Appendix S2.1).

For visualization purposes, we constructed ordination plots based on the negative binomial MM with three clusters (as our interest was to look for differences in terms of relative species composition; see discussion in Section 'Ordination via mixture modelling') and the best fitted LVM. We compared these plots with one based on nMDS, calculated using `metaMDS` in the `vegan` (Oksanen *et al.* 2013) package with Bray–Curtis distances and applying the Wisconsin double standardization, as well as one based on DCA using the `decorana` function in the `vegan` package. Similar patterns were observed between LVM, nMDS and, to a certain extent, DCA (Fig. 2a–c). Specifically, all three ordination plots suggested sites grouped into

approximately three clusters – sites {22–24, 26–28} (right), sites {8, 15–21} (top left), sites {1–7, 9–14} (bottom), with site 25 falling between these clusters. The MM plot shared some similarities with the other methods, but also some differences, with the most obvious being the separation of sites {24,27} from {22,23,26,28}.

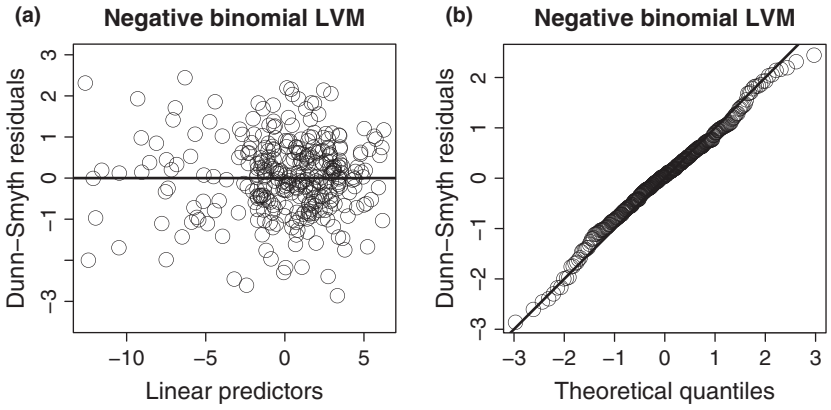

Fig. 1. Plots of Dunn–Smyth residuals vs. linear predictors and normal quantile plots of the residuals from the best fitting LVM (a and b), and the corresponding Poisson LVM (c and d). The best fitting LVM used a negative binomial model with site effects and two latent variables (Table 1). Note the Poisson LVM displays a fan-shaped pattern indicative of overdispersion, whereas the negative binomial LVM displays no clear pattern. Normal quantile plots also suggest that the negative binomial LVM better satisfies the residual normality assumption compared to the Poisson LVM.

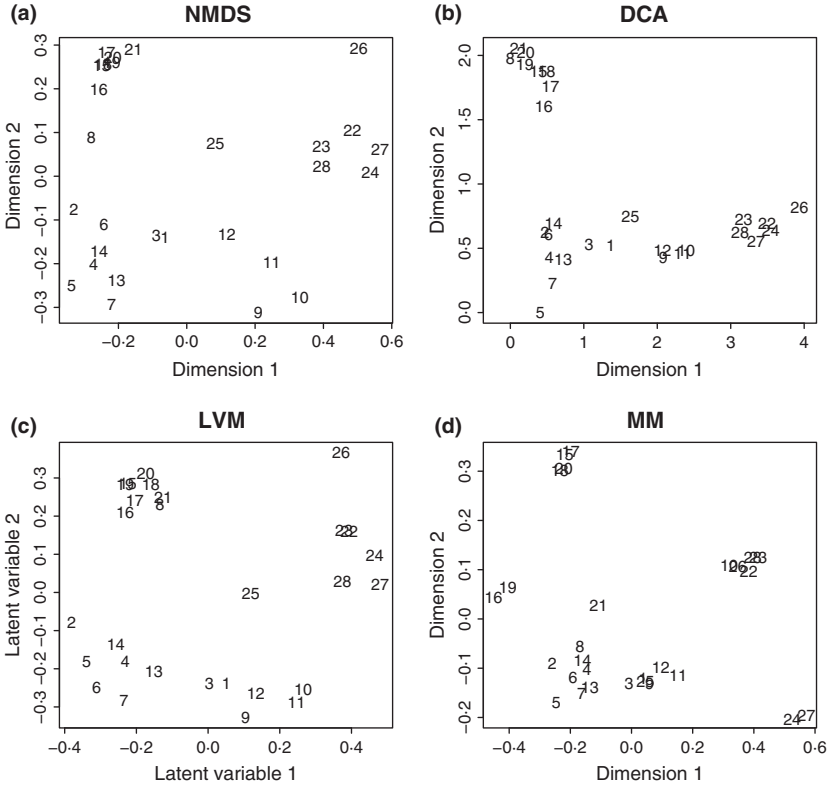

Fig. 2. Ordination plots for spider data set, obtained from: (a) nMDS; (b) DCA; (c) the best fitted LVM; and (d) a negative binomial MM with three clusters. Sites (row) numbers for the data set have been used as the symbols. Both model-based ordination plots have been rotated and rescaled using Procrustes transformation, to better highlight any similarities/differences with the nMDS plot.

Our focus in this study is the development of unconstrained ordination approaches, but this grouping of sites as observed in the nMDS and LVM plots suggest two directions to go in subsequent analyses – firstly, a mixture model which clustered both sites *and* species might have been more effective in capturing the site clustering in the data. Such a biclustered mixture model could be regarded as a model-based alternative to correspondence analysis and is discussed in Pledger & Arnold (2014). Secondly, environmental variables were measured at the 28 sites, and comparison to these using regression approaches suggested that the underlying gradient driving changes in species composition is well characterized by environmental variables. In particular, fitting a GLM of each species' counts against six environmental variables available (soil dryness, % bare sand, % fallen leaves, % cover moss, % cover herb layer, reflection of soil surface) accounted for 77% of the deviance in species composition explained by the best fitted LVM (see Appendix S2.1 for R code for how this value was calculated).

In summary, the unconstrained ordination discussed above provides a useful step in understanding the differences in community composition across sites. Use of a model-based approach instead of nMDS returned similar results, but offered a number of checks along the way regarding the distribution of data, the adequacy of a two-dimensional ordination (in fact it was the optimal choice for this data set), and whether data better conformed to a gradient model (LVM) or a species clustering model (MM) for compositional change.

CORAL DATA SET

The second data set consists of 10 transects of a coral assemblage located in the Tikus Islands, Indonesia (Warwick & Clarke 1990), with presence–absence data collected in 1981 and again in 1983. This is an example data set from the PRIMER manual (Clarke & Warwick 2001), which has been reanalysed on a number of occasions using nMDS and other distance-based approaches (Clarke 1993; Warwick & Clarke 1993; Anderson 2006). In each case, the conclusion was that there is a substantial change in dispersion between the two sampling times, although plots of the raw data strongly suggest otherwise (Warton 2008) and the apparent dispersion has since been shown to be an artefact of failure to model the mean–variance relationship correctly (Warton, Wright & Wang 2012).

For simplicity, we only considered species with more than four presences over the two years. We also removed one record for a site in 1983 that contained no presences, as it provided no information about the composition or relative abundance. It also causes some computational issues for most methods. The final data set contained $n = 19$ sites and $P = 18$ species. For this data set, it is important to note there was an El Niño event in 1982–1983 that led to substantial coral bleaching, causing a tenfold decrease in site total abundance between the two sampling times. This can be easily seen in the raw data, for instance using a comparative boxplot of the row sums by year (not shown).

Both the nMDS and DCA plots are suggestive of an increase in dispersion following the El Niño event (Fig. 3a and b). This was interpreted as evidence that stress increases variability in coral communities (Warwick & Clarke 1993). Using BIC, the best MM used one cluster and the best LVM used one latent variable without site effects (see Appendix S2.2). Residual analysis indicated both the best MM and LVM had good fits to the data (figures not shown). For the purposes of ordination (comparison with DCA and nMDS), we constructed two-dimensional scatter plots using MMs with two and three clusters, and LVMs with two latent variables. However, as discussed below, results from one-dimensional ordination plots led to the same conclusions as in the two-dimensional plots.

Model-based ordinations of relative species abundance (i.e. without site effects) displayed a strong location effect with no evidence of a difference in dispersion across the two sampling times (Fig. 3c and d). This strong location effect is strikingly clear on inspection of the raw data, yet such a pattern is surprisingly difficult to reproduce using nMDS – in numerous reanalyses of this data set, the only comparable ordination we are aware of is Fig. 6(c) in Anderson *et al.* (2011).

The MM- and LVM-based ordinations of species composition (i.e. controlling for site effects) led to an interesting additional result – they did not offer any substantial evidence of a change in composition between the two sampling times (Fig. 3e and f), with observations from 1981 and 1983 interspersed on the plot. Therefore, it seems the only discernible pattern from our ordinations is the strong bleaching of coral of all species.

Finally, note that although two-dimensional plots have been used here for consistency with nMDS and DCA, we could have in fact come to the same conclusions using an MM and LVM with one cluster and one latent variable, respectively. Indeed, the best fitted models based on BIC were those containing only one cluster or latent variable (see Appendix S2.2). One-dimensional ordination plots based on these models present very similar patterns to those observed in Fig. 3, that is the strong location effect for models without site effects and the lack of any pattern for models controlling for site effects.

Simulation studies

We undertook two types of simulation to demonstrate the efficiency of our method – data-driven simulations, as presented here, and some simulations using the COMPAS software, as in the next section.

In our data-driven simulation study, we generated data under each of the MM and LVM models, but in such a way that simulated data to mimic the properties of the two observed data sets analysed in the previous section. This led to four true models to simulate from the following: (i) a negative binomial LVM with two latent variables and site effects fitted to the spider data set, (ii) a negative binomial MM with three clusters fitted to the spider data set, (iii) a Bernoulli LVM with two latent variables and no site effects fitted to the coral data set and (iv) a Bernoulli MM with two clusters fitted to the coral data set. All four models were fitted as part of the worked examples in Section 'Worked examples', but we now

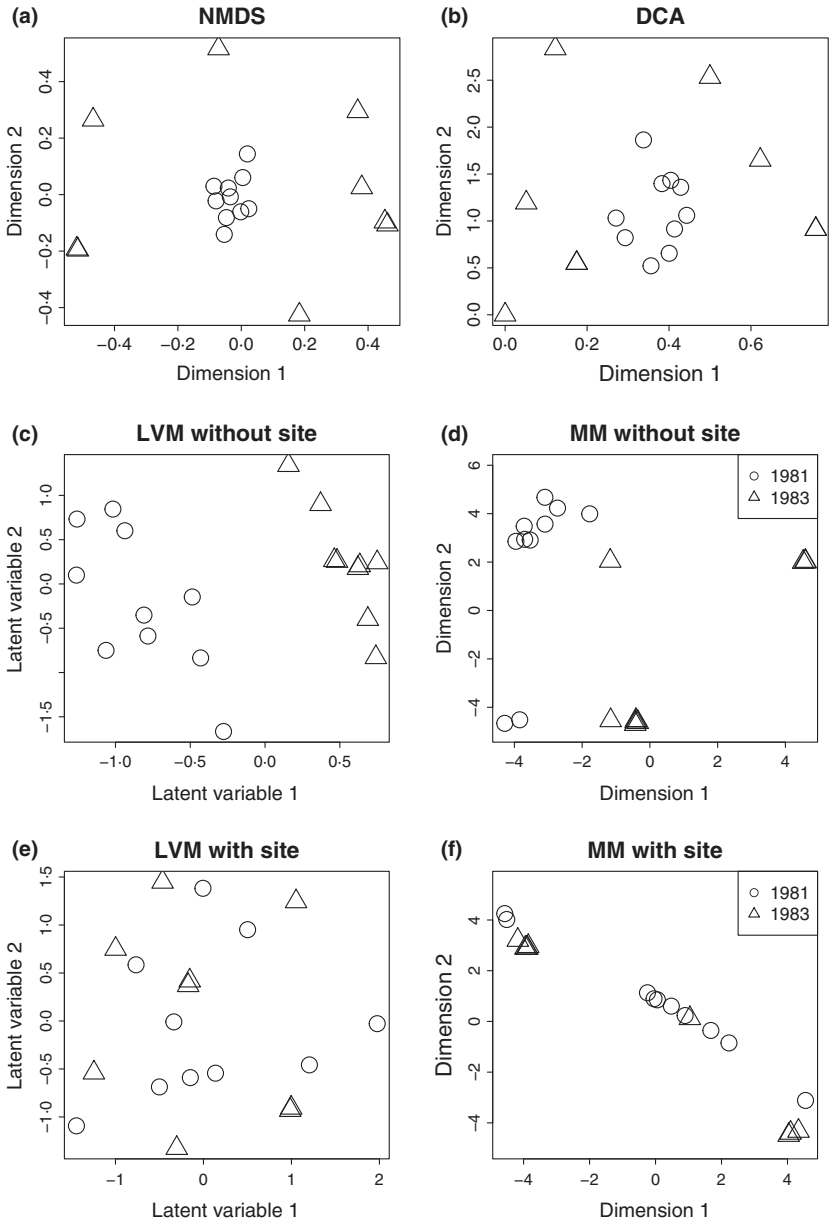

Fig. 3. Ordination plots for coral data set, obtained from: (a) nMDS; (b) DCA; (c) MM and (d) LVM without site effects; (e) LVM and (f) MM with site effects. The nMDS plot incorrectly exhibits a strong dispersion effect. Both the LVM and MM correctly identify the strong location effect (decrease in species richness) between the two sampling times. After adjusting for differing richness across sites and sampling times, the MM and LVM ordination plots show no any significant pattern indicating that, in fact, there is no evidence of a difference in species composition before and after the El Niño event.

use their estimated fits as a basis for simulating data that mimicked properties of the observed data. We simulated 200 data sets for each true model. For each data set, we compared the two model-based approaches to nMDS (using Bray–Curtis distance and applying the Wisconsin double standardization) and DCA, and assessed performance by computing the symmetric Procrustes error between the estimated and the true ordinations for each approach via the procrustes function in the `vegan` package. Briefly, Procrustes errors are calculated by applying a Procrustes transformation, involving some combination of uniform scaling (expansion and contraction) and rotation, to minimize the squared differences between the ordinations of the fitted model and the true ordinations,

$$\text{Procrustes Error} = \sum_{i=1}^{n}\sum_{r=1}^{q}(z_{ir,\text{fitted}} - z_{ir,\text{true}})^2,$$

where $z_{ir,\text{fitted}}$ denotes the Procrustes rotated ordination coordinate for site i and latent variable r from the fitted model (for MM, these are the elements of $\boldsymbol{\gamma}_i$), and $z_{ir,\text{true}}$ denotes the corresponding true ordination coordinate.

We also considered nMDS in conjunction with Hellinger and Chord dissimilarities, and with the Bray–Curtis dissimilarity applied to fourth-root transformed count data, as well as CA, with very similar results as those presented below (see Appendix S3). Template R code used to perform the simulations is provided in Appendix S5.

It is important to emphasize that the sole aim of this simulation study was to empirically present the concept of efficiency in model-based ordination (see Section 'Advantages of model-based approaches'). It is not meant to be an extensive comparison of model-based and algorithm-based approaches to ordination – such a comparison is as an avenue for future research.

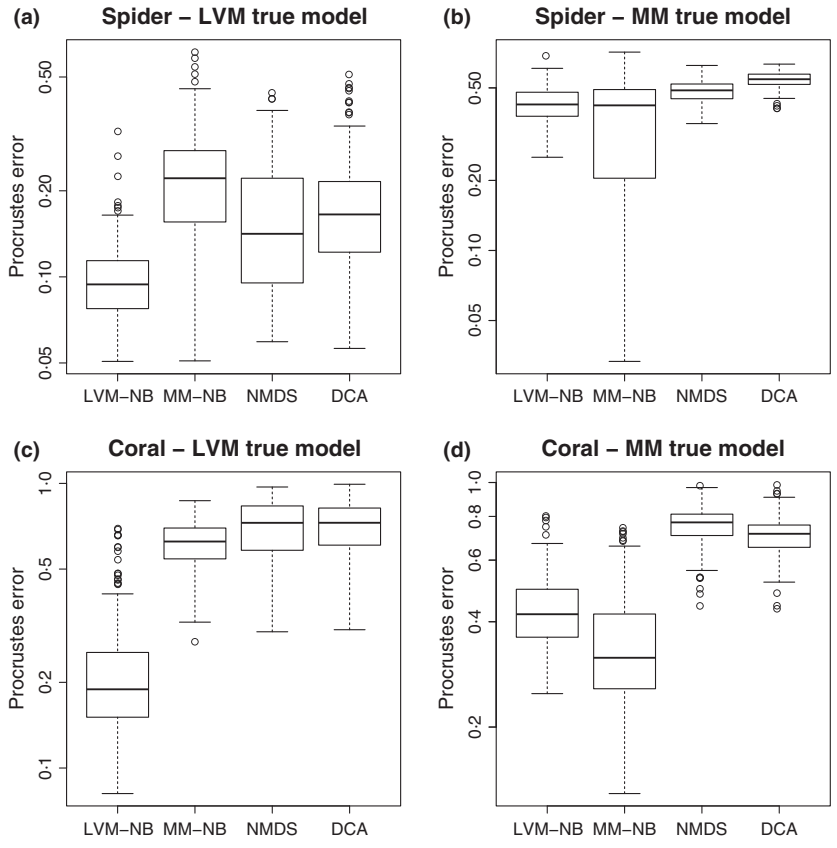

Fig. 4. Comparative boxplots of Procrustes errors between true and estimated ordination points. In plots a and b, the true models were negative binomial LVMs and MMs fitted to the spider data set, respectively, and we compared negative binomials MMs (MM-NB), negative binomial LVMs (LVM-NB), nMDS and DCA. In plots c and d, the true models were Bernoulli MM and LVMs fitted to the coral data set, respectively, and we compared Bernoulli MMs and LVMs, nMDS, and DCA. In all plots, it is evident that when the fitted model matches the true data-generation mechanism, model-based ordination methods perform better at recovering the true ordinations.

When the model used to analyse the data was close to true data-generation mechanism, using a model-based ordination more effectively recovered the true locations of sites along their underlying gradients (Fig. 4). This illustrates the concept of efficiency with model-based approaches to ordination – fitting an LVM (MM) to multivariate abundance data simulated from an LVM (MM) led to lowest Procrustes error of the methods compared. Furthermore, BIC almost always selected the correct model type (LVM or MM) to use in the four simulation settings. Not surprisingly, the associated Procrustes errors were also the lowest (or even slightly lower) of all the methods compared (see Appendix S5).

Additionally, these simulations provide some preliminary evidence that model-based approaches (especially LVMs) can still perform well when their assumptions are not satisfied. This result is most strongly evidenced when we consider the two settings where the true models were MMs – even though LVMs are the wrong model here, that is, they make the wrong assumptions about the data (see Section 'A comparison of the model-based methods'), the Procrustes error was still lower than both nMDS and DCA (Fig. 4b and d). Note also that even for the LVM simulations, the assumptions of the LVMs fitted were still not quite correct. Specifically, recall that LVMs assume the latent variables are drawn from independent, standard normal distributions; see below eqn (2). On the other hand, the true model ordinations on which Fig. 4(a and c) are based (as shown in Figs 2c and 3c, respectively) are far from being normally distributed.

SIMULATION FROM COMPAS

To further explore the idea model-based approaches may perform well even when their assumptions are not satisfied, we conducted an additional, small simulation study when generating data from a very different underlying model. Here we use COMPAS (Minchin 1987), a well-known programme for simulating multivariate abundance data of counts. Briefly, COMPAS generates unimodal curves representing species responses to an indirect (latent) gradient. Count responses are then generated based on a combination of turnover (beta-diversity), noise (mimicking sampling error for instance) and the addition of 'marginal/vagrant' species. We used an implementation of COMPAS available in the `CommEcol` package (Melo 2013).

We considered two sets of true site ordinations: Setting 1 contained) $n = 30$ sites, with sites {1–15} and {16–30} generated from bivariate Gaussian distributions with mean vectors (30,60) and (60,40), respectively. This produced two clusters of sites; Setting 2 contained $n = 45$ sites, with sites {1–15}, {16–30}, and {31–45} generated from bivariate Gaussian distributions with mean vectors (30,60), (60,40) and (80,60), respectively. This produced three clusters of sites. The true site coordinates for both sets are shown in Appendix S4. Note that, although the mean vectors are quite different, the clusters are not easily distinguishable because the covariance matrices used in the bivariate Gaussian distributions are fairly large (and different for each cluster). For both sets of true ordinations, we

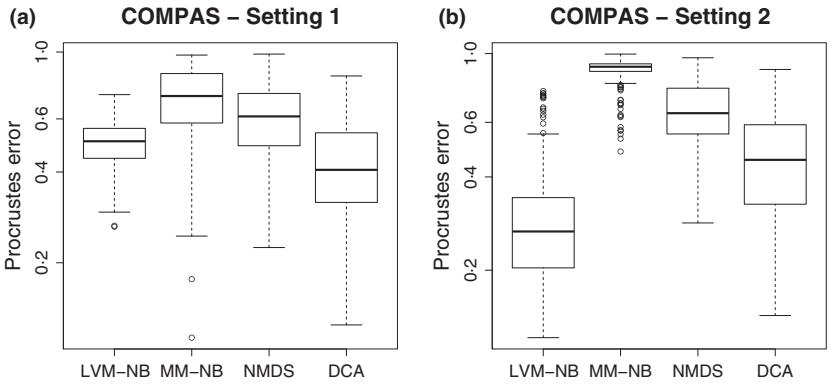

Fig. 5. Comparative boxplots of Procrustes errors between true and estimated ordination points for simulations based on COMPAS. Plot a is based on Setting 1, plot b is based on Setting 2 (with three clusters of sites). In Setting 1, DCA performed best followed (although it did exhibit substantial variability) by the negative binomial LVM, while in Setting 2 the negative binomial LVM significantly outperformed nMDS and DCA.

simulated count data sets with $p = 30$ species and used the values recommended in the `CommEcol` package for the other parameters in COMPAS, that is abundance of species at its modal point set to 2 (log scale) for both gradients, beta-diversity set to 2 for both gradients, qualitative noise set to 0.3, and 10% marginal species (see Melo 2013, for further details on parameter inputs). We simulated 500 data sets for both settings.

As COMPAS was designed to simulate differences between sites in terms of species composition, we included site effects in the LVMs and fitted MMs with three clusters. As in Section 'Simulation studies', we compared the model-based approach to nMDS (using Bray–Curtis distance and applying the Wisconsin double standardization) and DCA. As with the previous set of simulations, we also considered nMDS in conjunction with some other dissimilarity measures and transformations and CA, with results similar to those below (see Appendix S4).

In Setting 1, DCA has the lowest lower Procrustes errors followed by the negative binomial LVMs (Fig. 5a). In Setting 2, the model-based approach was the clear best performer, with its median Procrustes error more than half that of nMDS and roughly half that of DCA (Fig. 5b). While MM had weaker performance in both settings, the values of BIC were in fact lower for the negative binomial LVM compared to the MM in almost all the simulated data sets. In other words, had we used model selection in each simulated data set to choose between the two model-based ordination approaches, then we would have chosen an ordination method (LVM) whose performance was either close to or significantly better than algorithm-based approaches at capturing site locations along their underlying gradients.

Discussion

We have described two model-based approaches to unconstrained ordination, based on mixture models and latent variables models. These methods are based on well-known model-based dimension reduction techniques, and work by grouping species (mixture models) or by placing sites on an underlying gradient (latent variable models). Both offer a number of advantages to constructing ordinations, including formal methods of assumption checking, model selection and the ability to explicitly control for spurious data properties such as mean–variance relationships. Simulations show these methods lead to more reliable ordination plots if the fitted model is similar to the true data-generation mechanism, and are suggestive that these methods may lead to more reliable ordination plots even when assumptions are not satisfied (as in Fig. 5).

While MMs and LVMs were fitted by maximum likelihood in this study, this is by no means the only available method of estimation. Indeed, even within the maximum likelihood framework, there are many options for algorithms, a conspicuous alternative to our approach to fitting LVMs is adaptive quadrature (as used in GLLAMM software, Skrondal & Rabe-Hesketh 2004). Both MMs and LVMs could also be fitted using hierarchical Bayesian methods (Cressie *et al.* 2009), which would require the additional and sometimes non-trivial steps of prior specification, and if using simulation-based approaches, convergence diagnosis. Both maximum likelihood and Bayesian methods have the same advantages of model-based methods to ordination discussed above.

The model-based approaches proposed here are quite flexible and capable of handling nonlinear forms of species response to latent gradients, consistently outperforming nMDS in simulations (Fig. 5). In the case of LVMs, this arises despite the use of a linear model, with the site effects (α_i in eqn 2) able to capture a common form of nonlinearity across species (e.g. unimodality with a common tolerance). The latent variables themselves are also capable of handling nonlinearity, in a manner somewhat analogous to a recent result for generalized linear mixed models (Jamil & ter Braak 2013). However, if a particular type of nonlinearity is to be expected, then an obvious next step would be to adjust the model to incorporate this information. For example, to handle unimodal response with different tolerances across species, one could add quadratic terms to eqn (2), although our initial attempts on this front encountered numerical instability. A similar issue was encountered for presence/absence data by Walker & Jackson (2011), who addressed the issue using regularization. A potential alternative is to include a quadratic main effect term only, thus to handle unimodal responses with common tolerance explicitly, rather than indirectly via the α_i.

While the results from our small scale study in Section 'Simulation studies' showed very promising results in favour of model-based methods, a more comprehensive simulation study

would be welcome, to compare the model-based approaches to unconstrained ordination proposed above to algorithm-based methods currently in use. This remains an avenue of future research.

A current disadvantage of model-based ordination is computation time – LVMs can be quite time-consuming relative to nMDS in moderately sized data sets. However, the model-based methods have the advantage that computation time scales linearly with sample size, whereas nMDS is quadratic or slower. For example, while our LVM function under default settings takes noticeably longer for data sets with less than 100 observations (although rarely more than a few minutes), it can actually be faster than nMDS for data sets with thousands of observations (when both methods take a few hours). Currently, our code is in an early stage of development, and there are a number of ways computation time can be accelerated.

An advantage of model-based approaches to ordination not discussed prior to this point is the unified framework it offers, and the extensions that are possible. The methods described in this study are just a starting point and can be extended in a number of ways to answer a range of important ecological questions. For example, abundance data may not come as counts or presence–absence, and other choices of distributional assumptions may better handle biomass (Dunstan *et al*. 2013), ordinal data (Yee 2010) and per cent cover data (an area of future research, as they are currently no adequate distributional assumptions that can suitably handle the statistical properties of per cent cover data). Model-based ordination could equally well be applied outside of the context of multivariate abundance data – for example, to species traits data (see Moles *et al*. 2013, for a similar idea). Furthermore, both model-based methods introduced here can quite naturally be extended to construct biplots (Gabriel , 1971, 1998), for which the ordination provides additional insight into how species composition changes along the underlying gradient. Both MMs and LVMs can also incorporate environmental variables and could be used to devise model-based methods for constrained ordination. Identifying which environmental variables are important for each species then becomes subsumed in the model selection process.

As seen in the worked examples, the assumption of a standard normal distribution for the latent variables does not appear to impose strong restrictions on the pattern of sites in the ordination site. Nevertheless, it is possible to relax this normality assumption and consider other distributional assumptions, or even attempt to estimate the distribution of the latent variables in a nonparametric way (see discussion in Chapter 4 Skrondal & Rabe-Hesketh 2004).

One key distinction between an LVM (with environmental variables) and other model-based multivariate regressions (e.g. Wang *et al*. 2012; Jamil *et al*. 2013a) is that the LVM offers a parsimonious way to incorporate species interaction or other sources of species correlation (such as missing covariates). Marginalizing over latent variables in a q-dimensional LVM, the linear predictors at a site are multivariate normal with a rank-q matrix R of species correlations. The estimate of R is tractable even when the number of species p is not small, when using an unstructured, full rank matrix in modelling would not be an option. The main reason multivariate analysis in ecology tends to rely on resampling for inference is that no reliable model-based approach to inference has previously been available which can account for species correlation. Extensions of the models considered here may be a way to overcome this.

Acknowledgements

FKCH is supported by a Research Excellence Award from the University of New South Wales and a CSIRO PhD Scholarship. ST was supported by the Academy of Finland Grant 256291. SDF was supported by the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Research Program (NERP). DIW is supported by an Australian Research Council Future Fellowship. Thanks to Gerry Quinn and Jakub Stoklosa for useful discussions.

Data accessibility

Both data sets used in the worked examples are available from the R package mvabund (Wang *et al*. 2012), and are referred to as spider and tikus.

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

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

Anderson, M.J. (2006) Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, **62**, 245–253.

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

Anderson, M.J., Crist, T.O., Chase, J.M., Vellend, M., Inouye, B.D., Freestone, A.L. *et al*. (2011) Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecology Letters*, **14**, 19–28.

ter Braak, C.J. (1985) Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics*, **41**, 859–873.

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

ter Braak, C.J. & Prentice, I.C. (1988) A theory of gradient analysis. *Advances in Ecological Research*, **18**, 271–317.

Clarke, K.R. (1993) Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, **18**, 117–143.

Clarke, K.R. & Warwick, R.M. (2001) *PRIMER v5: User Manual/Tutorial*. PRIMER-E Limited, Plymouth, UK.

Cressie, N., Calder, C.A., Clark, J.S., Hoef, J.M.V. & Wikle, C.K. (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, **19**, 553–570.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B. Statistical Methodology*, **39**, 1–38.

Dunn, P.K. & Smyth, G.K. (1996) Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.

Dunstan, P., Foster, S., Hui, F. & Warton, D. (2013) Finite Mixture of Regression Modeling for high-dimensional count and biomass data in Ecology. *Journal of Agricultural, Biological and Environmental Sciences*, **18**, 357–375.

Faith, D.P., Minchin, P.R. & Belbin, L. (1987) Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, **69**, 57–68.

Gabriel, K.R. (1971) The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, **58**, 453–467.

Gabriel, K.R. (1998) Generalised bilinear regression. *Biometrika*, **85**, 689–700.

Gauch, H., Chase, G.B. & Whittaker, R.H. (1974) Ordination of vegetation samples by Gaussian species distributions. *Ecology*, **55**, 1382–1390.

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

Hill, M.O. (1974) Correspondence analysis: a neglected multivariate method. *Applied Statistics*, **23**, 340–354.
Hill, M.O. & Gauch Jr, H. (1980) Detrended correspondence analysis: an improved ordination technique. *Vegetatio*, **42**, 47–58.
Hui, F.K., Warton, D.I., Foster, S. & Dunstan, P. (2013) To mix or not to mix: comparing the predictive performance of mixture models versus separate species distribution models. *Ecology*, **94**, 1913–1919.
Ihm, P. & van Groenewoud, H. (1975) A multivariate ordering of vegetation data based on Gaussian type gradient response curves. *Journal of Ecology*, **63**, 767–777.
Ihm, P. & van Groenewoud, H. (1984) Correspondence analysis and Gaussian ordination. *COMPSTAT Lectures*, **3**, 5–60.
Jamil, T. & ter Braak, C.J. (2013) Generalized linear mixed models can detect unimodal species-environment relationships. *PeerJ*, **1**, e95.
Jamil, T., Ozinga, W.A., Kleyer, M. & ter Braak, C.J. (2013) Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *Journal of Vegetation Science*, **24**, 988–1000.
Knott, M. & Bartholomew, D.J. (1999) *Latent Variable Models and Factor Analysis. 7*. Edward Arnold, London, UK.
Kooijman, S. (1977) Species abundance with optimum relations to environmental factors. *Annals of Systems Research* (eds B. Van Rootselaar & H. Koppelaar), pp. 123–138. Springer, USA.
Kruskal, J.B. (1964a) Multidimensional scaling by optimizing goodness of fit to a non metric hypothesis. *Psychometrika*, **29**, 1–27.
Kruskal, J.B. (1964b) Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115–129.
Legendre, P. & Anderson, M.J. (1999) Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, **69**, 1–24.
Legendre, P. & Legendre, L. (2012) *Numerical Ecology, Volume 20*. Elsevier, Oxford, UK.
McCullagh, P. & Nelder, J. (1989) *Generalized Linear Models*. Chapman & Hall, London, UK.
McLachlan, G. & Peel, D. (2004) *Finite mixture models*. John Wiley & Sons, USA.
Melo, A.S. (2013) CommEcol: Community Ecology Analyses. R package version 1.5.9/r38.
Minchin, P.R. (1987) Simulation of multidimensional community patterns: towards a comprehensive model. *Vegetatio*, **71**, 145–156.
Moles, A.T., Peco, B., Wallis, I.R., Foley, W.J., Poore, A.G., Seabloom, E.W. *et al*. (2013) Correlations between physical and chemical defences in plants: tradeoffs, syndromes, or just many different ways to skin a herbivorous cat? *New Phytologist*, **198**, 252–263.
Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B. *et al*. (2013) vegan: Community Ecology Package. R package version 2.0-10.
Pledger, S. & Arnold, R. (2014) Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. *Computational Statistics & Data Analysis*, **71**, 241–261.
Quinn, G.G.P. & Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge, UK.
Scheffer, M. & van Nes, E.H. (2006) Self-organized similarity, the evolutionary emergence of groups of similar species. *Proceedings of the National Academy of Sciences, USA*, **103**, 6230–6235.
Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
Seong, T.-J. (1990) Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, **14**, 299–311.
Skrondal, A. & Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multi-level, Longitudinal and Structural Equation Models*. Chapman & Hall, Boca Raton.
Van der Aart, P. & Smeenk-Enserink, N. (1974) Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology*, **25**, 1–45.
Van der Vaart, A.W. (2000) *Asymptotic Statistics, Volume 3*. Cambridge University Press, Cambridge, UK.
Walker, S.C. & Jackson, D.A. (2011) Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. *Ecological Monographs*, **81**, 635–663.
Wang, Y., Naumann, U., Wright, S.T. & Warton, D.I. (2012) mvabund – an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, **3**, 471–474.
Warton, D.I. (2008) Raw data graphing: an informative but under-utilized tool for the analysis of multivariate abundances. *Austral Ecology*, **33**, 290–300.
Warton, D.I., Wright, S.T. & Wang, Y. (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, **3**, 89–101.
Warwick, R.M. & Clarke, K.R. (1993) Increased variability as a symptom of stress in marine communities. *Journal of Experimental Marine Biology and Ecology*, **172**, 215–226.
Warwick, R., Clarke, K.R. & Suharsono (1990) A statistical analysis of coral community responses to the 1982–83 El Nino in the Thousand Islands, Indonesia. *Coral Reefs*, **8**, 171–179.
Wedel, M. & Kamakura, W. (2001) Factor analysis with (mixed) observed and latent variables. *Psychometrika*, **66**, 515–530.
Yee, T.W. (2010) The VGAM package for categorical data analysis. *Journal of Statistical Software*, **32**, 1–34.

Received 9 June 2014; accepted 16 July 2014
Handling Editor: Robert B. O'Hara

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Theoretical details for latent variable models.

Appendix S2. Additional results for worked examples.

Appendix S3. Additional results for simulations.

Appendix S4. Additional results for COMPAS simulations.

Appendix S5. Example R code for performing the simulation studies.

Appendix S6. Example R code for fitting latent variable models.