CellPress

Review

So Many Variables: Joint Modeling in Community Ecology

David I. Warton,[1,*] F. Guillaume Blanchet,[2] Robert B. O'Hara,[3] Otso Ovaskainen,[4,5] Sara Taskinen,[6] Steven C. Walker,[2] and Francis K.C. Hui[7]

Technological advances have enabled a new class of multivariate models for ecology, with the potential now to specify a statistical model for abundances jointly across many taxa, to simultaneously explore interactions across taxa and the response of abundance to environmental variables. Joint models can be used for several purposes of interest to ecologists, including estimating patterns of residual correlation across taxa, ordination, multivariate inference about environmental effects and environment-by-trait interactions, accounting for missing predictors, and improving predictions in situations where one can leverage knowledge of some species to predict others. We demonstrate this by example and discuss recent computation tools and future directions.

A New Phase for Community Modeling in Ecology

Many of the questions posed in ecology require the consideration of **abundance** (see Glossary, including presence/absence) collected simultaneously across multiple taxonomic groups, for example species. The abundances in different taxa typically form the **response variables** in a **multivariate analysis** and are analyzed for several different goals, recent examples include: to study the impact of experimental removal of invasive crayfish on macroinvertebrate communities [1], to find taxa that can act as indicators of biodiversity loss due to logging and oil palm disturbance [2], to find leading environmental correlates of feral cat diet via meta-analysis [3], and to predict microbial interaction networks from co-occurrence data [4].

The number of taxa in an assemblage is typically larger than what can be modeled using classical multivariate analyses [5] – for example, the alpine plant data of Box 2 were sampled across 75 sites and represent almost as many species (65). If organisms are identified using modern tools such as DNA barcoding and metabarcoding [6,7], their number can be especially large, often in the thousands. Further, the data often have many zeros, and the **samples** therefore may not be rich in information – the European tree data analyzed in Box 3 had over 3000 plots, but half the species were each found in only 50 plots or less. Historically, the large number of taxa to be jointly analyzed, relative to the information available on each, has been technically challenging [8]. However, this is changing rapidly.

Analysis tends to follow one of two methodological traditions. The older tradition is 'algorithmic' [9] multivariate analysis, which in ecology has had a historic focus on algorithms for **ordination** (e.g., correspondence analysis, non-metric multidimensional scaling, canonical correspondence analysis) [5,10,11], and resampling-based hypothesis testing procedures [12,13]. While an underlying statistical model may sometimes have served as motivation [11], technology limited

Trends

Many ecological questions require the joint analysis of abundances collected simultaneously across many taxonomic groups, and, if organisms are identified using modern tools such as metabarcoding, their number can be in the thousands.

While historically such data have been analyzed using *ad hoc* algorithms, it is now possible to fully specify joint statistical models for abundance using multivariate extensions of generalized linear mixed models.

These modern 'joint modeling' approaches allow the study of correlation patterns across taxa, at the same time as studying environmental response, to tease the two apart.

Latent variable models are an especially exciting tool that has recently been used for ordination as well as for studying the factors driving co-occurrence.

[1]School of Mathematics and Statistics, and Evolution & Ecology Research Centre, The University of New South Wales (UNSW), Sydney, Australia
[2]Department of Mathematics and Statistics, McMaster University, Hamilton, Canada
[3]Biodiversity and Climate Research Centre, Frankfurt, Germany
[4]Metapopulation Research Center, Department of Biosciences, University of Helsinki, Finland
[5]Centre for Biodiversity Dynamics, Department of Biology, Norwegian

 http://dx.doi.org/10.1016/j.tree.2015.09.007

the extent to which the fitting algorithm could reflect it. There is a second and more recent tradition of species distribution modeling, in particular, methods described as community-level modeling [14–17], which focus more on predictive modeling and mapping the distribution of species and species diversity, with less focus on correlations across species.

We are now entering a third phase in methods for multivariate analysis in ecology. This has been driven by the advent of sophisticated hierarchical modeling tools, a watershed for complex problems that arise in ecology [18]. For the first time we can fully specify a joint statistical model for abundance across many taxa, and hence incorporate in a single model the impact on abundance of environmental predictors and interspecific interaction. This paper reviews these new methods, with a particular focus on the **latent variable model** (LVM) as a flexible tool which can address a range of analysis goals, including all those in the examples at the start of this section.

Joint Models for Abundance

The methods described in this paper are all extensions of the **generalized linear model** (GLM) [19], widely used to model abundance (e.g., [20–22]). A **joint model** necessarily requires the inclusion of random effects, hence some form of mixed model [23], to capture correlation in abundance across taxa. There are several ways to proceed, and a key issue to consider is the level of complexity in the model. A balance needs to be found between using a sufficiently simple model that its parameters can be estimated reliably from available information, and using a sufficiently complex model that it can realistically capture the main forms of correlation.

A simple way to incorporate correlation is to introduce it indirectly via a random univariate effect applied to each sample [24]. The presence of a common random effect across taxa in a sample induces a constant positive covariance between taxa. However, one would rarely expect covariance across all taxa to be constant or always positive.

A complicated way to incorporate correlation is to introduce it directly via a multivariate random effect applied to each sample, to form a multivariate **generalized linear mixed model** (GLMM, Box 1 and Figure 1, Key Figure) [25–28]. This model is especially useful when the number of taxa is small compared to the number of samples, and can be fitted using standard mixed modeling software (e.g., lme4, Box 4). A difficulty, however, is that the multivariate random effect typically is assumed to have a completely unstructured variance–covariance matrix. The number of parameters increases quickly as the number of taxa increases, presenting a problem for estimation and inference. For example, the alpine plant data of Box 2 represent 65 species, leading to a GLMM with over 2000 covariance parameters, most of which did not converge during model-fitting (Appendix B in the supplementary material online).

A flexible way to incorporate correlation is to use a LVM (Box 1 and Figure 1) [29,30] which introduces some unobserved ('latent') predictors to each sample. The latent variables induce correlation between taxa, and their number controls model complexity, such that it is possible to fit joint models across many taxa. This is a key advantage because the number of taxa is frequently large. LVMs have previously been used in contexts where the number of response variables is in the thousands, such as in the analysis of microarray data [31], made possible by substantially reducing the number of covariance parameters in the model. For example, the LVM fitted to the alpine plant data in Box 2 involved only 129 covariance parameters, all parameters successfully converged, and computation time was almost one tenth of that for the multivariate GLMM.

It is helpful to think of latent variables as resembling the axes in an ordination; in fact, an important use of LVMs is as a model-based approach to ordination [32,33]. The latent variables, similarly to

University of Science and Technology, Norway
[6]Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland
[7]Mathematical Sciences Institute, Australian National University, Canberra, Australia

*Correspondence:
david.warton@unsw.edu.au
(D.I. Warton).

Box 1. What is a Joint Model for Abundance?

A joint model for abundance (y_{ij}, for sample $i = 1, \ldots, n$ and taxon $j = 1, \ldots, m$) describes correlation across taxa as well as response to measured predictors ($\mathbf{x}_i$), as in Figure 1 in main text. Below are two especially useful hierarchical approaches to specifying a joint model.

Multivariate Generalized Linear Mixed Model (GLMM)

The GLMM approach extends the generalized linear model (GLM) [19] by specifying multivariate random effects u_{ij} for each sampling unit to capture correlation across taxa [25,27,28]. Mean abundance m_{ij} could be assumed to be:

$$g(m_{ij}) = \alpha_i + \beta_{0j} + \mathbf{x}_i' \beta_j + u_{ij} \quad [I]$$

where $g(\cdot)$ is the **link function**, $\mathbf{x}'$ is the transpose of vector $\mathbf{x}$, and for each taxon j, β_{0j} is an intercept and β_j is a vector of regression coefficients related to measured predictors. The site effect α_i is optional and adjusts for site total abundance or richness [33], to focus on modeling relative abundance or composition rather than absolute abundance.

This model is hierarchical because the $\mathbf{u}_i = (u_{i1}, \ldots, u_{im})$ as well as the abundance y_{ij} are treated as random:

$$\begin{array}{ll} y_{ij}|\mathbf{u}_i & \sim F(m_{ij}, \phi_j) \\ \mathbf{u}_i & \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \end{array} \quad [II]$$

where '$\sim$' means 'is distributed as', and $F(m_{ij}, \phi_j)$ is the assumed distribution of y_{ij} characterized by its mean m_{ij} and possibly a dispersion parameter ϕ_j – typically Gaussian for continuous data, binomial for presence/absence, and Poisson or negative binomial for counts. The variance-covariance matrix of random effects $\boldsymbol{\Sigma}$ controls the correlation between taxa, and is assumed to be completely unstructured (apart from constraints on variances needed for presence/absence data [29]). This part of the model is problematic when the number of taxa (m) is large because the number of parameters in $\boldsymbol{\Sigma}$ increases rapidly (quadratically) with m.

Latent Variable Model (LVM)

A LVM is a function of unmeasured predictors (or 'latent variables'), $\mathbf{z}_i$, as well as measured predictors. One way to specify the LVM is:

$$\begin{array}{ll} g(m_{ij}) & = \alpha_i + \beta_{0j} + \mathbf{x}_i' \beta_j + u_{ij} \\ \text{where } u_{ij} & = \mathbf{z}_i' \lambda_j \end{array} \quad [III]$$

the important distinction being that u_{ij} must now be linearly related to a set of latent variables $\mathbf{z}_i$. Estimation is difficult because neither the latent variables $\mathbf{z}_i$ nor the factor loadings λ_j are known. For estimation to be possible, either prior assumptions are needed [39] or additional constraints required (Appendix A in the supplementary material online). A dispersion term is sometimes included (termed a 'specific factor' in factor analysis) such that overdispersion can be modeled separately from the covariance modeling afforded by the latent variable term.

The latent variables $\mathbf{z}_i$ are treated as random (to account for the fact that they are unobserved) as well as the abundance y_{ij}, commonly by assuming:

$$\begin{array}{ll} y_{ij}|\mathbf{z}_i & \sim F(m_{ij}, \phi_j) \\ \mathbf{z}_i & \sim N(\mathbf{0}, \mathbf{I}) \end{array} \quad [IV]$$

This can be understood as a type of multivariate GLMM with the constraint $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$, where $\boldsymbol{\Lambda}$ is the full matrix of factor loadings, with the λ_j as its columns. This has potentially many fewer parameters than the previous GLMM because $\boldsymbol{\Lambda}$ only has as many columns as there are latent variables, while an unstructured $\boldsymbol{\Sigma}$ has as many columns of parameters as there are taxa.

ordination axes, have the dual purposes of representing missing predictors and of representing the main axes of (co)variation of abundance across taxa. Historically, ordination axes have been interpreted as latent variables [5,10,11,34,35]. What is different within the LVM framework, as compared to previous ordination algorithms, is that the notion of latent variables is made explicit in the statistical model (Equation III in Box 1), and these latent variables are treated as random effects. Treating the unobserved latent variables as random is useful for technical reasons, and also provides the capacity to directly model correlations between taxa.

Latent variables have been described as 'interaction currencies' [26], in other words quantities that mediate interactions between taxa but whose measured values are not included in the model. This is similar to thinking of latent variables as missing predictors, but it puts the emphasis on the idea that correlations across taxa are induced by these missing predictors. LVMs were

Glossary

Abundance: the extent to which a type of organism is present in a sample unit, measured either as a count, biomass, % cover, a factor with ordered levels, or presence/absence.

Continuous variable: a variable that can take any value within some interval (cf. discrete variable). Abundance is rarely continuous, complicating the modeling process.

Discrete variable: a variable that can take one of a countable number of distinct values. Abundance is often discrete, for example counts could be 0, 1, 2, 3,...

Credible interval (95%): a range of values with 95% (posterior) probability for a parameter (a Bayesian version of a 95% confidence interval).

Generalized linear model (GLM): a regression model to predict a response variable assuming it comes from a distribution in the exponential family (Poisson, binomial, etc.), and assuming that some known transformation of the mean response is a linear function of predictor variables.

Generalized linear mixed model (GLMM): a GLM with random effects included: that is, some of the coefficients are assumed to come at random from a larger population of potential values [23]. Of particular interest here is where the random effect is multivariate to account for correlation (Box 1).

Joint model, or joint species distribution model: a parametric statistical model for the abundance of multiple taxa (usually species), accounting for correlation between taxa as well as response to predictor variables.

Latent variable model (LVM): a regression model for multivariate data that includes some unobserved ('latent') predictors that are usually introduced to model correlation or to account for missing predictors (Box 1).

Link function: the function in a GLM defining the transformation of the mean to a linear function of predictors (e.g., logit or probit for presence/absence data, log for counts). Its main purpose is to map from the scale of the linear predictor (which can take any real-numbered value) onto the scale on which the mean response is defined (e.g., all positive numbers for abundance).

Key Figure

Schematic Diagram of Joint Models for Response of Abundance to Predictors, and Residual Correlation Across Taxa

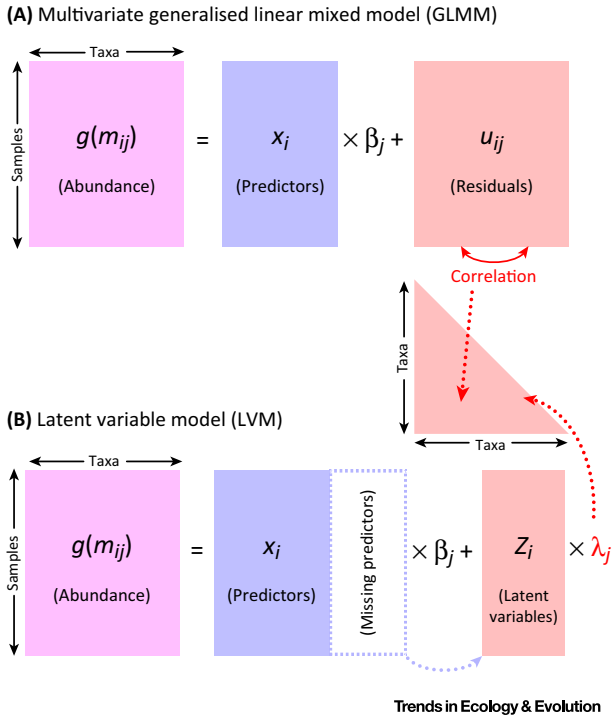

Figure 1. Correlation can be handled in different ways: (A) A multivariate GLMM uses correlated multivariate random effects, u_{ij}, to estimate correlation. (B) A LVM includes a smaller number of latent variables, $\mathbf{z}_i$, which play the role of missing predictors. Their factor loadings, λ_j, approximate the correlation across taxa, but use fewer parameters than the GLMM (because the matrix of $\mathbf{z}_i$ has fewer columns than the matrix of u_{ij}).

Multivariate analysis: joint analysis of multiple response variables; in particular, a joint analysis for abundance of multiple taxonomic groups.
Ordination: a visualization tool attempting to represent the main structures in multivariate data together with a reduced set of usually two or three axes.
Predictor variable: a variable used to predict the response of interest. In this paper these are treatments, environmental variables, or functional traits.
Residual correlation: correlation between response variables that is not explained by predictors in the model. Joint models can estimate this.
Response variable: variable of primary interest in analysis (e.g., for which predictions are required). In this paper these are typically abundances in taxa.
Sample: the sampling unit at which abundances are measured for all taxa, often a site or transect.

recently used very much for this purpose [36], finding that negative interactions between shrub species at fine-scales could largely be explained by hydrological variables.

The most common example of a latent variable model is factor analysis, and LVMs are sometimes even called factor analytic models [30]. Factor analysis has been used in ecology for over 60 years [34], largely as a tool for estimating the underlying causes of covariation in a set of response variables [5]. In Box 1 we show how the method generalizes to include measured

Box 2. Analyzing Response of Alpine Plants to Snowmelt Date

Snowmelt dates have become earlier in many parts of the world in response to climate change [70], and it is of interest to understand the potential impact on plant communities. We will consider alpine plants at 75 sites in Aravo (Valloire), south east France [71], with varying snowmelt dates. Presence/absence records of 65 species are used in the analysis, each species having more than four presences. The data are available from the R package ade4 [72].

To study how (and why) plant communities vary along the snowmelt gradient, we fitted a latent variable model to predict the (probit of) probability of an observed presence of species j in site i as a quadratic function of mean snowmelt date (Julian day, averaged over 1997–1999), specific leaf area (SLA), and two latent variables:

$$\Phi^{-1}(m_{ij}) = a_i + \beta_{0j} + \text{snow}_i\beta_1 + \text{snow}_i^2\beta_2 + \text{snow}_i\text{SLA}_j\beta_3 + \text{snow}_i b_j + \mathbf{z}_i'\lambda_j, \quad [I]$$

where a_i and b_j are random effects that behave as error terms, soaking up cross-site and cross-species variation (respectively) that is not explained by the effects of snowmelt date and its interaction with SLA. This is a type of 'fourth

corner model' [24,64], predicting abundance as a function of environmental variables and functional traits, β_3 being the coefficient characterizing this interaction. The model was fitted using Bayesian MCMC estimation via JAGS [73] (see Appendix E in the supplementary material online for code), although similar models can also be fitted using the HMSC package (Box 4).

This model can be used to look at a few aspects of the problem:

Estimating species correlation: the residual correlation matrix (Figure I) suggests some fairly strong correlations between species even after controlling for the effects of snow melt and SLA (e.g., the entry in the first row of Figure I is 0.83).

Unconstrained and partial (residual) ordination: correlations can be visualized in a biplot, using an unconstrained ordination (i.e., no predictors, Figure IIA) or a 'partial ordination' [11] of residuals (Figure IIB) after controlling for effects of snowmelt, SLA, and their interaction. Sites have a gradient in snowmelt date, from the bottom right to the top left on the unconstrained ordination (Figure IIA), while species coordinates (black squares) suggest two sets of indicator species which preferred earlier and later snowmelt dates. Most of this gradient is removed on inclusion of snowmelt in the model, along with much of the cross-site variation (Figure IIB).

Multivariate inference: the β_3 term in the model is a fourth corner coefficient [64] that attempts to explain interspecific variation in response to snowmelt date using SLA, because SLA is a key trait characterizing plant ecological strategy that can vary across climate gradients [74]. The posterior mean estimate for β_3 was 1.21, with a 95% credible interval of (0.56, 1.73), suggesting strong evidence that plant species with higher SLA were more likely to be found in late-melting habitats.

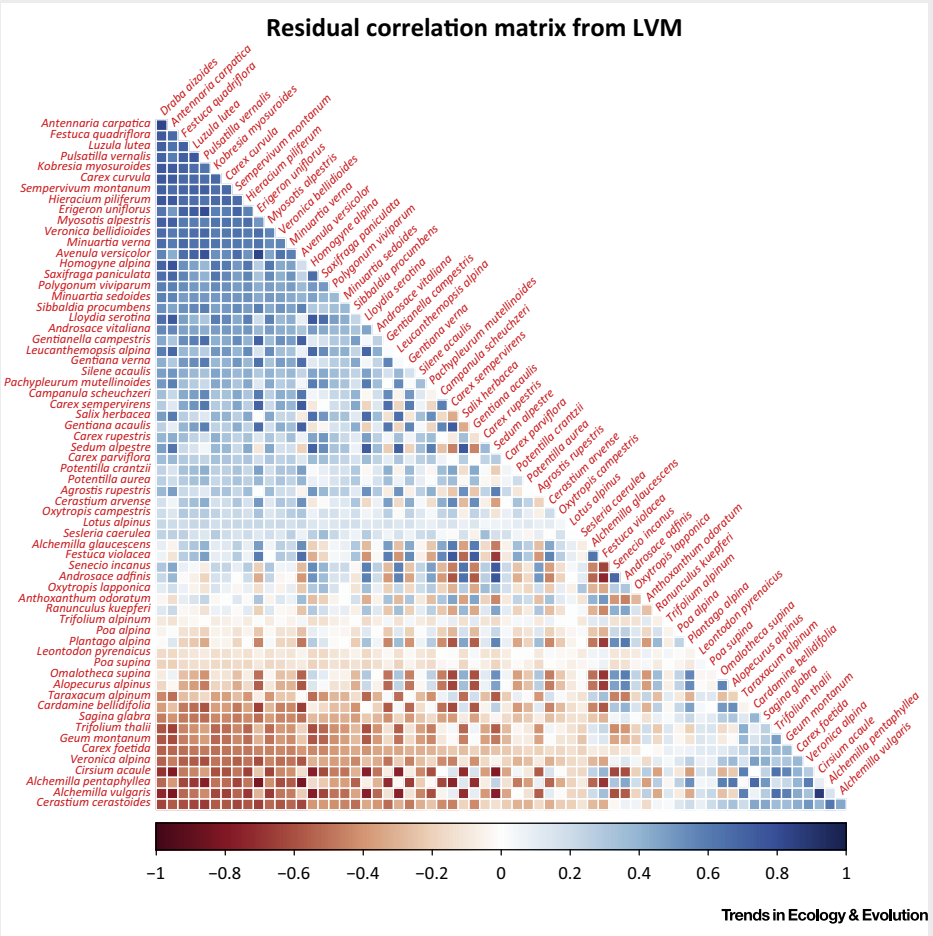

Figure I. Color Plot of the Residual Correlation Matrix, Based on Fitting the LVM Specified in Equation I to the Alpine Plant Data. Species have been ordered by factor score, such that positively correlated species tend to be close together.

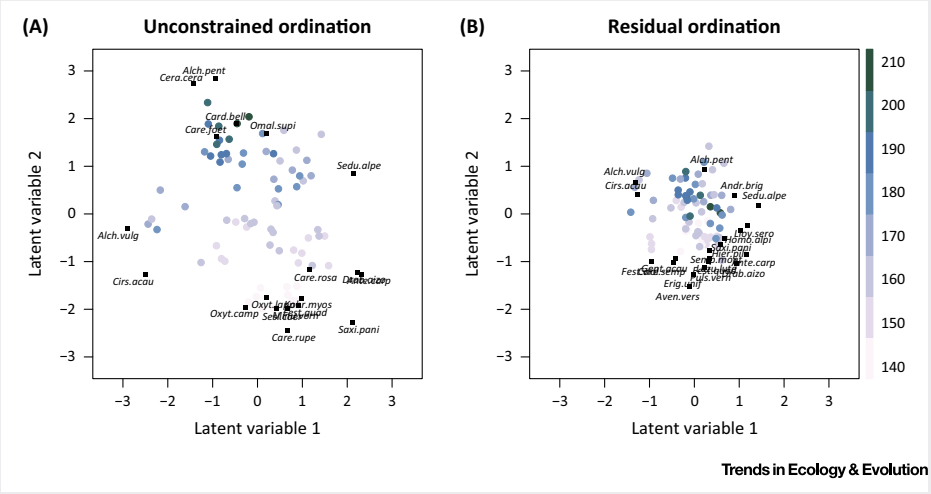

Figure II. Model-Based Biplots of the Alpine Plant Data, (A) for the unconstrained model, (B) after including snowmelt and SLA as predictors in Equation I. Sites are shown in colors indexed by the value of snowmelt date, and the 20 species with the largest factor loadings (in terms of distance from the origin) in each analysis are shown as black squares. Species in the same direction and far from the origin are highly correlated, for example *Alchemilla pentaphyllea* and *Cerastium cerastoides* in (A), and there are fewer such species after controlling for snowmelt date as in (B).

predictors and to handle non-normal responses. Both these extensions are necessary to model abundances across many taxa as a function of environmental variables. Adding measured predictors to a factor analysis has seen some application in ecology under the guise of structural equation modeling [37], for example, to tease apart the relative roles of ecological and energetic constraints as determinants of species richness [38]. Handling non-normal responses has been a major technical challenge demanding modern computational techniques for estimation [29], and, as such, ecologists have only recently begun to explore this approach to joint modeling in community ecology [32,33].

An issue to consider when fitting latent variable models (as in factor analysis) is the choice of the number of latent variables. This controls model complexity – less latent variables means a simpler model, but at the risk of a poorer approximation of the true correlation structure, whereas more latent variables means a more complex model closer to a multivariate GLMM. This trade-off can be managed by standard model selection approaches such as information criteria [33], or (in a Bayesian framework) by using a model that automatically shrinks less-informative latent variables to zero [39]. We have found that the step-up in model complexity on adding latent variables is fairly steep and a small number (less than five) is often sufficient for a good approximation to correlations.

The hierarchical approaches in Box 1 are the most common ways to specify a joint model for abundance, but other strategies are possible. The generalized estimating equations (GEE) approach [40] can and has been used for a similar purpose [41], although it is best suited to situations where the correlation is treated as a nuisance rather than being of interest in itself [42]. Another class of models with potential are copulas, widely used in finance [43] and risk analysis [44], where the key idea is to couple quantiles from a marginal model of response variables (e.g., a GLM) with quantiles from a multivariate model of correlation across responses. Copulas have made few inroads into ecology to date (although see [45]), in part because flexible approaches that can handle discrete data have only recently been developed [46,47].

In the same way as there are alternatives to hierarchical models for specifying a joint model, there are alternatives to latent variables for modeling correlations across taxa. For example, graphical models assume that correlations are driven by a smaller network of key species interactions [48,49]. The phylogenetic GLMM models correlations as a function of phylogenetic relatedness [50,51], although including latent variables as well might be advisable, to detect correlations not related to phylogeny, as in model V of [50].

Equations I and III in Box 1 can be extended in several ways. One important way is to add predictors on taxa as well as on samples, as in Box 2. Another important extension is to treat the parameters used to model taxa $(\beta_{0j}, \boldsymbol{\beta}_j)$ as random rather than fixed. This might be useful if the taxa observed in a dataset are a sample from a larger pool of potential taxa, and we want to make inferences across this broader pool, for example [24].

While the focus here is on modeling abundance, the models discussed here can readily be used with other types of response variable; for example, to model demographic quantities such as tree mortality [52] or phenology [53]. A particular issue analyzing abundance data is that typically the response is not a **continuous variable** and therefore cannot be modeled under the assumption of multivariate normality, and this complicates the modeling process.

Joint models have several exciting applications in ecology, spanning the study of interactions between taxa, predictive modeling, descriptive tools, confirmatory analyses, and accounting for missing predictors. Some key applications are discussed below.

Modeling Residual Correlation Between Taxa

An important application of joint models is in estimating the correlation between taxa that arises for reasons not attributable to the measured predictors included in the model. Such correlation could be due to biotic interactions such as competition and facilitation, although the exact type of biotic interaction cannot be inferred from co-occurrence [4,54]. It could also be due to joint response to unmeasured predictors, or to other forms of misspecification of the mean model [26].

If the number of taxa is small, correlations could be estimated directly via a GLMM with a multivariate random effect [27,28] as in Box 1. In other situations, a LVM could be used as a more parsimonious method of modeling correlation.

Correlation in a LVM is controlled by the factor loadings (Box 1), which are easily converted into a correlation matrix characterizing **residual correlation** across taxa. For example, in Box 2 (Figure I) a LVM with two latent variables was used to estimate the correlations between 65 alpine plant species, after controlling for the effects of snowmelt and specific leaf area (SLA).

By modeling response of abundance to predictors jointly with the correlation across taxa, we have a means of teasing the two apart. For example, it is possible to compare the amount of (co) variation before and after inclusion of predictors in the model, for example by using the trace of the variance–covariance matrix $\boldsymbol{\Sigma}$ [55]. In the alpine plant data (Box 2), the inclusion of snowmelt, SLA, and their interaction reduced the trace from 703 to 330; on this basis we could conclude that this combination of trait and environment predictors explains approximately 53% of the covariation in alpine plant species. One could also consider which environmental predictors best explain covariance across species, as in [36].

Model-Based Ordination

By treating latent variables as ordination axes, a LVM (commonly with two latent variables) can be understood as a model-based approach to unconstrained ordination [32,33]. A model-based

approach to ordination offers several advantages over traditional ordination methods. For example, models can be used to account for important (and otherwise spurious) data properties such as the mean–variance relationship [56]. Model selection and residual analysis tools can be used to verify key aspects of a model, such as the choice of the number of latent variables to consider and distributional assumptions [33]. Preliminary simulations [33] found that LVMs were better able to estimate the true location of sites along underlying gradients, as compared to distance-based ordination techniques such as non-metric multidimensional scaling ('nMDS') [57], even when assumptions of the model were not all met (with non-linear response to gradients and violations of distributional assumptions). Other advantages of using LVMs, in common with other model-based approaches in general, include the capacity to use them for predictive purposes and improved interpretability [9].

While the estimated values of latent variables $\mathbf{z}_i$ provide an ordination of sites, the factor loadings λ_j provide an ordination of taxa. For principal component analysis and correspondence analysis, both elements can be combined in a biplot, as in Figure II in Box 2.

While [32] and [33] proposed unconstrained model-based ordination, a novel extension is to construct an ordination after including any measured predictors thought to be important in characterizing the differences in abundance across taxa and samples. This leads to model-based 'partial ordination' [11] (as in Figure IIB in Box 2) which may be used (for instance) to study the extent to which the co-occurrence patterns suggested in an unconstrained ordination can be explained by shared response to environmental variables, or to provide a low-dimensional representation of these interactions across taxa. Unconstrained and partial ordinations can be presented on the same scale to give a visual indication of the reduction in amount of covariation after accounting for predictors, as in Box 2.

Multivariate Inferences about Predictors

Joint models, whether GLMMs or LVMs, can be used to make multivariate inferences about the effect of the **predictor variables** $\mathbf{x}_i$, while accounting for any residual correlation between taxa. Accounting for correlation between taxa, and doing so in a flexible way, is important to ensure that inferences made jointly across multiple taxa are statistically valid. Two examples of this are when studying how well species traits explain interspecific variation in environmental response (Box 2) and when predicting spatial patterns in species richness, with associated measures of uncertainty (Box 3).

When there are many taxa, currently the only viable alternative approach for making valid multivariate inferences is to resample sites to make design-based inferences about environmental effects [58]. Resampling, however, introduces its own challenges – it is computationally intensive, and incorporating random effects or accounting for spatiotemporal autocorrelation is difficult. LVMs are a fully model-based alternative with much potential in this regard.

We argued above that models with many taxa involve many parameters, demanding efforts to simplify the model to ensure that its components are all estimable from the data at hand. The same is true when we look only at the part of the model describing the effects of predictor variables. For example, in Box 2 there are 65 alpine species, and therefore if a separate quadratic snowmelt term were included for each species, this would add 130 parameters to the model only for this single predictor variable. One way forward is to make connections across taxa in their response to predictors, expecting that similar taxa will respond to their environment in a similar way. This can be done via reduced rank regression [59], by assuming that regression coefficients across taxa are drawn from a common distribution [60], by clustering them using a mixture model [61–63], or by using functional traits to predict variation across taxa in their response to predictor variables ('the fourth corner problem') [12,24,64]. Using functional traits, if

Box 3. Predicting Species Richness of European Trees

Our aim is to map tree species richness across Europe and study the uncertainty attached to such maps, using presence/absence records from 51 species collected as part of an international monitoring network [International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests (ICP Forest)] across 6118 forest plots [75]. The plots were regularly distributed across Europe, in a 16 × 16 kilometer grid covering over 2 million km^2 of forest, as in [76]. The distribution of each species (specifically, the probit of its probability of presence) was modeled as a quadratic function of growing degree days and summer precipitation, and species richness at a site (with a 95% credible interval) was estimated as the sum of predicted probabilities of presence across species. A latent variable modelLVM was used to account for correlation across species, an important consideration when estimating the uncertainty around species richness predictions, as in Figure I.

We used model validation on test data to study the effect on predictive performance of including latent variables, and the effect of the number of latent variables. Data were separated into training and test sets in a checkerboard pattern, the checkerboard being constructed using latitude and longitudinal lines every 5°. Alternating squares were assigned to training and test data, producing a total of 3134 training sites and 2984 test sites. Models with different numbers of latent variables were fitted to the training data, then used to predict species richness (with 95% credible intervals) at each test site. Models were fitted using boral (Box 4), with extensions to estimate uncertainty in species richness predictions (see Appendix E in the supplementary material online for code). The predicted values at test sites were compared to 'true' species richness, calculated as the prevalence at each of the test sites, $\sum_{j=1}^{m} y_{ij}$.

If correlation across species was ignored, in other words if there were no latent variables in the model, the 95% credible intervals were too narrow and did not capture the observed value for species richness sufficiently often (Table I). By contrast, LVMs, which accounted for correlation across species, had wider interval widths, closer to 95% coverage. Including more than one latent variable in the model had little effect, suggesting that the covariation across the 51 species was sufficiently well accounted for by one latent variable (at least, for the purposes of predicting species richness).

This analysis assumed that abundances were independent across samples, but future analyses could look at whether there is spatial structuring in abundances beyond that explained by predictors in the model by incorporating spatially autocorrelated latent variables, along the lines of [77].

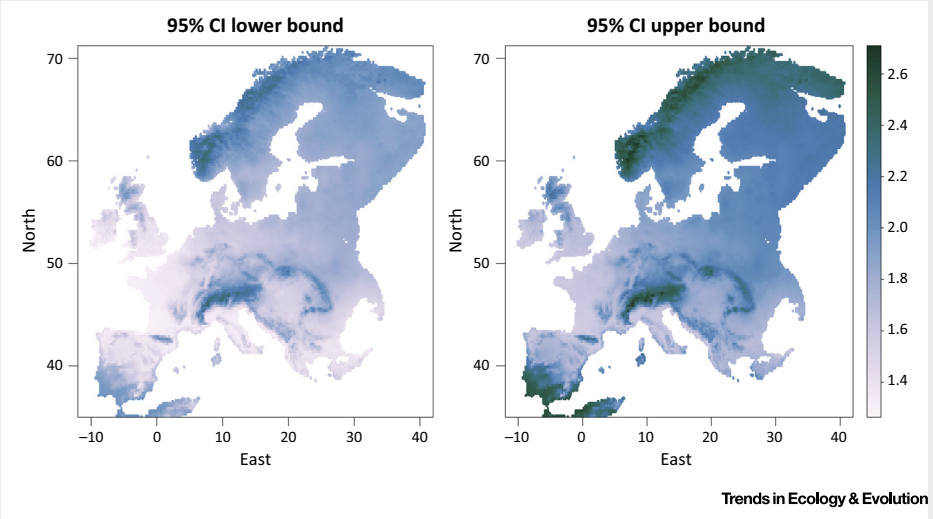

Figure I. Upper and Lower Limits of 95% Credible Intervals (CI) for Species Richness.

Table I. Predictive Success and Interval Width of 95% Credible Interval Predicting Species Richness at Test Sites, for the European Tree Data

Number of latent variables	0	1	2	3	4	5
Predictive success (%)	88.1	94.7	96.6	96.6	96.3	96.3
Mean interval width	4.66	6.13	6.32	6.50	6.57	6.55

available, is perhaps the simplest and most attractive of these approaches because it shifts our focus towards finding a functional explanation [65] for variation in response across taxa, as in Box 2.

Accounting for Missing Predictors

While diagnostic tools can be used to check assumptions, one can never be sure that all assumptions in the mean model are correct, and some violations remain hard to detect. One or more important predictors could be missing from the study, or perhaps the form in which the measured predictors enter the model is incorrect (e.g., assuming a quadratic response when the true response is more complex). The statistical term for such failures is 'misspecification' of the mean model [66]. Fortunately, univariate models fitted to carefully designed studies typically have some robustness to misspecification, with the predominant effect being to increase the amount of residual error, which can be readily accounted for (e.g., by modeling counts using negative binomial regression instead of Poisson). In a multivariate model, a secondary, important effect of misspecification of the mean model is that it induces residual correlation between taxa. This problem also arises in spatial modeling, where a missing predictor can induce spatial autocorrelation [67].

Using a multivariate GLMM/LVM can address the problem of misspecification because the random intercepts/latent variables play the role of missing predictors and enable inference despite misspecification of the mean model [68]. To illustrate, we conducted a small simulation study to look at the effect of omitting a predictor from the alpine plant analysis of Box 2 and, in particular, at the ability of 95% **credible intervals** for the fourth corner coefficient to capture its true value as specified in the simulation (see Appendix C in the supplementary material online for design and full results). A naïve model with no multivariate random effects nor latent variables had poor performance, returning biased estimates of the fourth corner coefficient and intervals that were too short, capturing the true value only 63% of the time. By contrast, the addition of a single latent variable was sufficient to remove bias and capture the true coefficient at a rate indistinguishable from the nominal level (94.4%).

Improving Predictions

When predicting abundance across a set of correlated taxa, joint models could improve predictive performance even if the model were correctly specified.

Joint models have a particular advantage for in-sample prediction because they can make use of correlations across taxa, which contain information useful for predicting abundance of one taxon from others. For example, when using a LVM, if predictions are made on the same samples that were used to fit the joint model, then one can condition on the estimated latent variables to construct more accurate predictions as well as narrower credible (confidence) intervals. This is analogous to predicting composition conditional on the location of a sample on an ordination. Alternatively, if there is a focal taxon in the community, then more accurate predictions for this particular taxon could be obtained by conditioning on its covariation with other taxa, as captured by the latent variables.

Returning to the problem of mapping species richness of European trees (Box 3), consider now predicting species richness and constructing 95% credible intervals back onto the training sites (rather than to independent test sites). If we condition on a single latent variable when making predictions, credible intervals for species richness can be reduced on average to 33% of their original width (Appendix D in the supplementary material online for further details). The scale of the reduction in interval width can be interpreted as an indication of

exactly how much uncertainty in predictions results from missing variables and/or covariation among species.

Another potential advantage, available to any model of multiple taxa (whether it accounts for correlation or not), is improved predictions as compared to single-taxon models by exploiting structure across taxa in their mean models. Improved prediction has been demonstrated when clustering environmental responses by taxon [62], assuming taxon-specific coefficients are drawn from a common distribution [25], or using functional traits to explain differences across taxa [64], as in Box 2. Incorporating such techniques into joint models has exciting potential.

Box 4. Fitting Joint Models

While there are several software packages that can fit a multivariate GLMM or a LVM, the fitting process is computationally challenging. This is largely because of the need to account for uncertainty in latent variables (or random effects) and the many correlation parameters to be estimated. The most common approaches (as usual) are maximum likelihood [78] and Bayesian estimation [79]. For maximum likelihood estimation, the main difficulty is integrating over unmeasured values, which usually need to be approximated numerically. For Bayesian estimation, the main difficulty is simultaneously estimating posterior distributions for a large number of parameters, and this can lead to convergence problems or demand long chains. Table I summarizes common strategies for addressing these difficulties.

The Laplace approximation is the most common maximum likelihood technique for estimating GLMMs and LVMs (Table I) – it makes simplifying assumptions that lead to a relatively fast estimation process, but which can be inaccurate in small samples [80]. Adaptive (Gauss–Hermite) quadrature improves accuracy but at some computational cost [81], especially when more than two latent variables are needed. Variational approximation [82] is a newer method that may strike a better balance, and LVM applications of this technique are under development. Expectation/maximization (EM) techniques are difficult to classify because their speed and accuracy depend on how they are implemented, for example [33] used a Monte Carlo implementation of the EM algorithm in which the size of the Monte Carlo sample trades off speed against accuracy. Markov chain Monte Carlo (MCMC) algorithms as in [83] use a similar trade-off and are the leading method of implementation for Bayesian models. BUGS [84], JAGS [73], and Stan [85] or INLA [86] software can be used to build Bayesian GLMMs or LVMs from the ground up (as in Appendix E in the supplementary material online), but special-purpose software (Table II) is available for most standard models.

The gllamm package on Stata [87] currently has the most functionality, but Stata is proprietary software and gllamm does not have an easy-to-use interface. Other packages in Table II are available on R. The ltm package, like gllamm, was originally written for the social sciences and handles binary or ordinal data. The lme4 package fits a multivariate GLMM relatively easily, but is harder to use for a LVM. Several packages were written specifically for fitting LVMs in ecology, and their main distinguishing features are that boral currently handles the broadest range of data types, HMSC is the most flexible in terms of types of model that can be fit, and mistnet is a predictive modeling tool that uses neural nets to fit nonlinear responses to predictors and latent variables. Starter code to fit joint models using lme4 and boral is available in Appendix E in the supplementary material online.

Table I. Methods for Estimating Joint Models

	Description	Speed[a]	Accuracy[a]	Software (Refs)
Laplace approximation	Approximate integrand as multivariate normal	***	*	lme4, gllamm
Adaptive quadrature	Approximate as a sum of 'optimal' function evaluations	**	**	gllamm
Variational approximation	Approximate the posterior as normal	***	**	[88]
Expectation/maximization (EM)	Re-express problem in a simpler 'complete data' form	*	***[b]	mistnet, ltm [33]
Markov chain Monte Carlo (MCMC)	Iteratively simulate missing values and parameters	*	***	boral, HMSC [83]

[a]Key: *, moderate; **, good; ***, excellent
[b]Depending on how it is implemented – if a Monte Carlo EM, accuracy and speed depend on the Monte Carlo sampler used.

Table II. Software Properties and Functionality

	gllamm	lme4	ltm	boral	HMSC	mistnet
Ref.	[29]	[89]	[90]	[91]	[92]	[68]
Free?	No	Yes	Yes	Yes	Yes	Yes
Multivariate GLMM?	Yes	Yes	No	No	No	No
LVM?	Yes	Yes[a]	Yes	Yes	Yes	Yes
Estimate number of LVs?	No	No	No	No	Yes	No
Ordinal data?	Yes	No	Yes[b]	Yes	No	Yes[a]
Counts?	Yes	Yes	No	Yes	Yes	Yes
Trait predictors?	Yes	Yes	Yes	No	Yes	Yes
Use for ordination?	Yes	Yes[a]	Yes	Yes	Yes	Yes
Inference?	Yes	Yes	Yes	Yes	Yes	No
Non-linear response?	No	No	No	No	No	Yes
Speed rating[c]	*	**	***	*	**	**

[a]This is not straightforward to implement (for lme4, using a modular approach as in [89]).
[b]Only for models with one latent variable and no measured predictors.
[c]Key: *, moderate; **, good; ***, excellent.

Concluding Remarks

Joint models are flexible tools with exciting potential for application in ecology, especially community ecology, where the number of taxa is rarely small compared to the number of samples. In such instances a latent variable approach can be used for a range of purposes, as discussed here, although this list is by no means exhaustive.

Both multivariate GLMMs and LVMs can be understood as special types of mixed effects models designed for multivariate data. Hence they can be used for much the same purposes as the use of random effects models in univariate analyses. Accounting for missing predictors, for example, is a common use for random effects in univariate modeling [69]. We have illustrated how joint models can be used to account for missing predictors in a multivariate setting. Other common uses of random effects are to account for random factors in a hierarchical sampling design, and to partition variance between different levels in the sampling hierarchy, and multivariate GLMMs or LVMs could be used for these same purposes in a multivariate context.

Computational techniques for joint models in community ecology are still evolving, especially for LVMs, but tools are already available to ecologists that can be used 'off-the-shelf' for the purposes described here, as in Box 4. However, the development of these models in ecology is still at an early stage. It is thus reasonable to expect advances in computational speed, accuracy, and ease-of-use as this branch of research matures.

There are several desirable extensions and outstanding issues (see Outstanding Questions), and one especially important issue is how to make multivariate inferences when the number of parameters in the mean model is large. Otherwise, joint models can only be used for multivariate inference about the effect of predictors when the number of taxa is small or when the response is assumed to be mediated by traits. Relaxing these constraints should be a priority for future methodological research in joint modeling.

Acknowledgments

D.I.W was supported by an Australian Research Council Future Fellowship (FT120100501) and an Australian Academy of Science travel grant. B.O'H. was supported by a LOEWE (Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer

Outstanding Questions

To what extent do environmental predictors explain co-occurrence? A joint model can answer this question directly, and applying this idea across different study systems could improve our understanding of the extent of co-occurrence through shared environmental response.

Do correlations between taxa vary spatially or temporally? It is commonly thought that species interactions will vary spatially and temporally, and a GLMM or LVM could be generalized to handle this. However, detecting changes in covariance patterns over space and time requires high-quality data.

Are there better ways to fit joint models? Latent variables are not the only tool that could be used to simplify the covariance model, and other options should be explored, as well as improved computation tools to fit joint models more efficiently. This is especially important for situations where there are many (operational) taxa, for example in metagenomics.

What extensions of joint models are needed? The basic joint models described here can be extended in several ways, including to estimate phylogenetic attraction or repulsion, to model non-linearity in response of abundance to latent variables, or to handle spatially or temporally correlated data, and further developments in these directions are needed.

Are the taxa we observe really all there is? Undetected species need to be considered if we wish to make inferences about the larger pool of taxa that could have been sampled, instead of conditioning on the taxa already observed. This may require use of random effects on taxa, and imperfect detection or missing data techniques, presenting new modeling challenges.

How to reduce the number of parameters in the mean model? To handle situations with many taxa, methods are necessary to reduce the number of parameters in the model for mean abundance, in much the same way as a LVM can reduce the number of parameters in the model of covariance. Possibilities include reduced rank regression, random effects, the LASSO or another sparse penalty, or mixture models, but applied in combination with latent variable models.

Exzellenz) initiative of the Hessian Ministry for Science and the Arts. O.O. and S.T. were supported by Academy of Finland grants 250444 and 251965, respectively. F.K.C.H. was supported by Australian Research Council discovery project grant DP140101259. We thank the editor, Cajo ter Braak, two anonymous reviewers, and the UNSW Eco-Stats group for helpful feedback that improved the manuscript, and Aidin Niamir for the European raster files.

Supplementary Information

Supplementary information associated with this article can be found online at http://dx.doi.org/10.1016/j.tree.2015.09.007.

References

1. Moorhouse, T.P. *et al.* (2014) Intensive removal of signal crayfish (pacifastacus leniusculus) from rivers increases numbers and taxon richness of macroinvertebrate species. *Ecol. Evol.* 4, 494–504
2. Edwards, D.P. *et al.* (2014) Selective-logging and oil palm: multitaxon impacts, biodiversity indicators, and trade-offs for conservation planning. *Ecol. Appl.* 24, 2029–2049
3. Doherty, T.S. *et al.* (2015) A continental-scale analysis of feral cat diet in australia. *J. Biogeogr.* 42, 964–975
4. Faust, K. and Raes, J. (2012) Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* 10, 538–550
5. Legendre, P. and Legendre, L. (2012) *Numerical Ecology.* (3rd edn), Elsevier
6. Valentini, A. *et al.* (2009) DNA barcoding for ecologists. *Trends Ecol. Evol.* 24, 110–117
7. Cristescu, M.E. (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends Ecol. Evol.* 29, 566–571
8. Warton, D.I. (2008) Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J. Am. Stat. Assoc.* 103, 340–349
9. Warton, D.I. *et al.* (2015) Model-based thinking for community ecology. *Plant Ecol.* 216, 669–682
10. Gauch, H. (1982) *Multivariate Analysis in Community Ecology,* Cambridge University Press
11. ter Braak, C.J.F. and Prentice, I.C. (1988) A theory of gradient analysis. *Adv. Ecol. Res.* 18, 271–317
12. Legendre, P. *et al.* (1997) Relating behavior to habitat: solutions to the fourth-corner problem. *Ecology* 78, 547–562
13. Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Aust. Ecol.* 26, 32–46
14. Ferrier, S. and Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *J. Appl. Ecol.* 43, 393–404
15. Elith, J. and Leathwick, J. (2007) Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Divers. Distrib.* 13, 265–275
16. Guisan, A. and Rahbek, C. (2011) SESAM: a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *J. Biogeogr.* 38, 1433–1444
17. Leitão, P.J. *et al.* (2015) Mapping beta diversity from space: Sparse generalised dissimilarity modelling (SGDM) for analysing high-dimensional data. *Methods Ecol. Evol.* 7, 764–771
18. Cressie, N. *et al.* (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecol. Appl.* 19, 553–570
19. Dobson, A.J. and Barnett, A. (2008) *An Introduction to Generalized Linear Models,* CRC Press
20. O'Hara, R.B. and Kotze, D.J. (2010) Do not log-transform count data. *Methods Ecol. Evol.* 1, 118–122
21. Warton, D.I. and Hui, F.K.C. (2011) The arcsine is asinine: the analysis of proportions in ecology. *Ecology* 92, 3–10
22. Szöcs, E. and Schäfer, R.B. (2015) Ecotoxicology is not normal. *Environ. Sci. Pollut. Res.* 22, 13990–13999
23. Bolker, B.M. *et al.* (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* 24, 127–135
24. Jamil, T. *et al.* (2013) Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *J. Vegetation Sci.* 24, 988–1000
25. Ovaskainen, O. *et al.* (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology* 91, 2514–2521
26. Kissling, W.D. *et al.* (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *J. Biogeogr.* 39, 2163–2178
27. Pollock, L.J. *et al.* (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol. Evol.* 5, 397–406
28. Clark, J.S. *et al.* (2014) More than the sum of the parts: forest climate response from Joint Species Distribution Models. *Ecol. Appl.* 24, 990–999
29. Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models,* CRC Press
30. Bartholomew, D.J. *et al.* (2011) *Latent Variable Models and Factor Analysis: A Unified Approach,* Wiley
31. Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.* 2, 418–427
32. Walker, S.C. and Jackson, D.A. (2011) Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. *Ecol. Monogr.* 81, 635–663
33. Hui, F.K.C. *et al.* (2015) Model-based approaches to unconstrained ordination. *Methods Ecol. Evol.* 6, 399–411
34. Goodall, D. (1954) Objective methods for the classification of vegetation. iii. an essay in the use of factor analysis. *Aust. J. Bot.* 2, 304–324
35. Jongman, R.H.G. *et al.* (1995) *Data Analysis in Community and Landscape Ecology,* Cambridge University Press
36. Letten, A.D. *et al.* (2015) Fine-scale hydrological niche differentiation through the lens of multi-species co-occurrence models. *J. Ecol.* (in press)
37. Grace, J.B. (2006) *Structural Equation Modeling and Natural Systems,* Cambridge University Press
38. Belmaker, J. and Jetz, W. (2015) Relative roles of ecological and energetic constraints, diversification rates and region history on global species richness gradients. *Ecol. Lett.* 18, 563–571
39. Bhattacharya, A. and Dunson, D.B. (2011) Sparse Bayesian infinite factor models. *Biometrika* 98, 291–306
40. Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22
41. Warton, D.I. (2011) Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics* 67, 116–123
42. Hardin, J.W. (2005) *Generalized Estimating Equations (GEE),* Wiley Online Library
43. Cherubini, U. *et al.* (2004) *Copula Methods in Finance,* John Wiley & Sons
44. Jongman, B. *et al.* (2014) Increasing stress on disaster-risk finance due to large floods. *Nat. Climate Change* 4, 264–268

45. de Valpine, P. *et al.* (2014) The importance of individual developmental variation in stage-structured population models. *Ecol. Lett.* 17, 1026–1038
46. Murray, J.S. *et al.* (2013) Bayesian Gaussian copula factor models for mixed data. *J. Am. Stat. Assoc.* 108, 656–665
47. Panagiotelis, A. *et al.* (2012) Pair copula constructions for multivariate discrete data. *J. Am. Stat. Assoc.* 107, 1063–1072
48. Friedman, J. *et al.* (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441
49. Abegaz, F. and Wit, E. (2015) Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Statistica Neerlandica* (in press)
50. Ives, A.R. and Helmus, M.R. (2011) Generalized linear mixed models for phylogenetic analyses of community structure. *Ecol. Monogr.* 81, 511–525
51. Kaldhusdal, A. *et al.* (2015) Spatio-phylogenetic multispecies distribution models. *Methods Ecol. Evol.* 6, 187–197
52. Clark, J.S. *et al.* (2014) Competition-interaction landscapes for the joint response of forests to climate change. *Global Change Biol.* 20, 1979–1991
53. Ovaskainen, O. *et al.* (2013) Community-level phenological response to climate change. *Proc. Natl. Acad. Sci. U.S.A.* 110, 13434–13439
54. Morales-Castilla, I. *et al.* (2015) Inferring biotic interactions from proxies. *Trends Ecol. Evol.* 30, 347–356
55. Legendre, P. *et al.* (2005) Analyzing beta diversity: Partitioning the spatial variation of community composition data. *Ecol. Monogr.* 75, 435–450
56. Warton, D.I. *et al.* (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods Ecol. Evol.* 3, 89–101
57. Kruskal, J.B. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1–27
58. Wang, Y. *et al.* (2012) Mvabund – an R package for model-based analysis of multivariate abundance data. *Methods Ecol. Evol.* 3, 471–474
59. Yee, T.W. and Hastie, T.J. (2003) Reduced-rank vector generalized linear models. *Stat. Model.* 3, 15–41
60. Ovaskainen, O. and Soininen, J. (2011) Making more out of sparse data: hierarchical modeling of species communities. *Ecology* 92, 289–295
61. Dunstan, P.K. *et al.* (2011) Model based grouping of species across environmental gradients. *Ecol. Model.* 222, 955–963
62. Hui, F.K.C. *et al.* (2013) To mix or not to mix: comparing the predictive performance of mixture models versus separate species distribution models. *Ecology* 94, 1913–1919
63. Dunstan, P.K. *et al.* (2013) Finite mixture of regression modelling for high-dimensional count and biomass data in Ecology. *J. Agric. Biol. Environ. Stat.* 18, 357–375
64. Brown, A.M. *et al.* (2014) The fourth-corner solution – using predictive models to understand how species traits interact with the environment. *Methods Ecol. Evol.* 5, 344–352
65. McGill, B.J. *et al.* (2006) Rebuilding community ecology from functional traits. *Trends Ecol. Evol.* 21, 178–185
66. Weisberg, S. (2014) *Applied Linear Regression.* (4th edn), Wiley
67. Dormann, C.F. *et al.* (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30, 609–628
68. Harris, D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods Ecol. Evol.* 6, 465–473
69. Wu, L. (2011) *Mixed Effects Models for Complex Data,* Chapman & Hall
70. Adam, J.C. *et al.* (2009) Implications of global climate change for snowmelt hydrology in the twenty-first century. *Hydrological Processes* 23, 962–972
71. Choler, P. (2005) Consistent shifts in alpine plant traits along a mesotopographical gradient. *Arct. Antarctic Alp. Res.* 37, 444–453
72. Dray, S. and Dufour, A. (2007) The ade4 package: Implementing the duality diagram for ecologists. *J. Stat. Software* 22, 1–20
73. Plummer, M. *et al.* (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* (Hornik, K. *et al.*, eds), pp. 20–22 Technische Universität Wien
74. Westoby, M. *et al.* (2002) Plant ecological strategies: some leading dimensions of variation between species. *Annu. Rev. Ecol. Syst.* 33, 125–159
75. Fischer, R. *et al.* (2010) *The Condition of Forests in Europe. 2010 Executive Report,* ICP Forests
76. Meier, E.S. *et al.* (2012) Climate, competition and connectivity affect future migration and ranges of European trees. *Global Ecol. Biogeogr.* 21, 164–178
77. Thorson, J.T. *et al.* (2015) Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods Ecol. Evol.* 6, 627–637
78. Millar, R.B. (2011) *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB,* John Wiley & Sons
79. Gelman, A. *et al.* (2013) *Bayesian Data Analysis,* CRC press
80. Huber, P. *et al.* (2004) Estimation of generalized linear latent variable models. *J. R. Stat. Soc. B* 66, 893–908
81. Rabe-Hesketh, S. *et al.* (2002) Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata J.* 2, 1–21
82. Ormerod, J. and Wand, M. (2010) Explaining variational approximations. *Am. Statistician* 64, 140–153
83. Walker, S.C. (2015) Indirect gradient analysis by Markov-chain Monte Carlo. *Plant Ecol.* 216, 697–708
84. Lunn, D.J. *et al.* (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* 10, 325–337
85. Stan Development Team. (2015) *Stan: A C++ Library for Probability and Sampling, Version 2.7.0*. http://mc-stan.org/
86. Rue, H. *et al.* (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. B* 71, 319–392
87. Stata Corporation (2001) *Stata Statistical Software Release 7.0: Programming,* Stata Corporation
88. Ormerod, J. and Wand, M. (2012) Gaussian variational approximate inference for generalized linear mixed models. *J. Comput. Graphical Stat.* 21, 2–17
89. Bates, D. *et al.* (2014) Fitting linear mixed-effects models using lme4. *arXiv 1406.5823*
90. Rizopoulos, D. (2006) ltm: an R package for latent variable modelling and item response theory analyses. *J. Stat. Software* 17, 1–25
91. Hui, F. (2015) *Package Boral (Version 0.9)*. Published online August 22, 2015. https://cran.r-project.org/web/packages/boral/boral.pdf
92. Blanchet, F.G. (2013) *HMSC: Hierarchical Modelling of Species Community*. Published online 18 April, 2013. http://rpackages.ianhowson.com/rforge/HMSC/man/HMSC-package.html