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).