Methods in Ecology and Evolution
British Ecological Society

Methods in Ecology and Evolution 2016, **7,** 744–750 doi: 10.1111/2041-210X.12514

APPLICATION

BORAL – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R

Francis K.C. Hui*

Mathematical Sciences Institute, The Australian National University, Canberra, ACT 0200, Australia

Summary

1. Model-based methods have emerged as a powerful approach for analysing multivariate abundance data in community ecology. Key applications include model-based ordination, modelling the various sources of correlations across species, and making inferences while accounting for these between species correlations.

2. BORAL (version 0.9.1, licence GPL-2) is an R package available on CRAN for model-based analysis of multivariate abundance data, with estimation performed using Bayesian Markov chain Monte Carlo methods. A key feature of the BORAL package is the ability to incorporate latent variables as a parsimonious method of modelling between species correlation.

3. Pure latent variable models offer a model-based approach to unconstrained ordination, for visualizing sites and the indicator species characterizing them on a low-dimensional plot.

4. Correlated response models consist of fitting generalized linear models to each species, while including latent variables to account for residual correlation between species, for example, due to unmeasured covariates.

Key-words: Bayesian inference, community composition, generalized linear models, hierarchical models, latent variable model, species interaction

Introduction

Many questions in ecology involve inference at a community-level rather than on individual species. For instance, we may be interested in how species composition varies across sites, and whether these differences can be explained by environmental covariates, treatments and so on. Multivariate abundance data are often collected to perform such community-level inference. Such data are characterized by a $n \times p$ matrix $\boldsymbol{Y}$, where rows $i = 1,\ldots,n$ are sites and columns $j = 1,\ldots,p$ are species recorded at each site.

With the vast improvements in computing power and statistical methodology over the past two decades, model-based methods have emerged as a powerful approach for analysing multivariate abundance data (e.g.Wang *et al*. 2012; Dunstan *et al*. 2013). Model-based methods provide a data-generating process and consequently a likelihood function that can be tailored to match the ecological processes and questions of interest. This is in contrast to distance-based methods, which focus more on describing the multivariate geometry of the data. In particular, hierarchical Bayesian models offer a framework for integrating many ecological processes into a single model (see Clark & Gelfand 2006; Royle & Dorazio 2008; Halstead *et al*. 2012, for discussions of the advantages of adopting a hierarchical Bayesian approach). Recent research has also shown that model-based approaches offer some advantages over distance-based methods for community-level inference, including the ability to straightforwardly check assumptions made in the analysis, more accurate predictions of the (in-sample) ordinations as well as species compositions and quantities such as species richness at new sites, and use of different measures for comparing models. Given the focus of this article is on application of new software, detailed discussion of these advantages are outside the scope of this work, and we refer the interested reader to Hui *et al*. (2014), Walker (2015), and Warton *et al*. (2015).

While the methodology underlying hierarchical models has progressed at great pace, the software remains underdeveloped, with only a handful of R packages currently available to ecologists for fitting models to multivariate abundance data. These include the BAYESCOMM package (Golding & Harris 2015), which is limited to analysing presence–absence data and does not come with methods for plotting ordinations, the MISNET (Harris 2015) package that uses machine learning techniques to build species assemblages for improved prediction, and the HMSC package (Blanchet 2014), which is in a developmental stage but aims to offer a general framework for fitting hierarchical Bayesian models. Most of the time however, ecologists are required to learn and write their own code in programs such as BUGS (Lunn *et al*. 2000) or JAGS (Just Another Gibbs Sampler, Plummer *et al*. 2003).

In this article, we introduce BORAL, an R package available on CRAN for Bayesian analysis of multivariate abundance data in ecology. Building upon the work of Hui *et al*. (2015), who proposed latent variable models for model-based unconstrained ordination, the BORAL package allows ecologists to fit

*Correspondence author: E-mail: fhui28@gmail.com a variety of models using Bayesian Markov chain Monte Carlo (MCMC) estimation. A key feature of the BORAL package is the ability to incorporate latent variables as a parsimonious method of modelling correlation between species. Without covariates, we can fit pure latent variable models in which species are regressed against a set of unknown covariates, leading to unconstrained ordination for visualizing site and species patterns. With covariates, we can fit correlated response models that combine separate species generalized linear models with latent variables to account for residual correlation, for example due to biotic interactions, missing covariates. This facilitates community-level inference about environmental covariates and treatments while controlling for extraneous correlation between species, as well as providing a method of 'residual ordination'; see Warton et al. (2015) for methodological details on how latent variable models are used for analysing multivariate abundance data). Note that the BORAL package relies on the JAGS program to perform the MCMC sampling. Users are therefore required to download and install JAGS separately from R, before installing the BORAL package.

We now list several main features of the BORAL package, which are illustrated in the worked examples to follow.

- A variety of distributions are available for modelling the species responses (columns of $\boldsymbol{Y}$). These include the normal distribution for continuous data, the Poisson and negative binomial distributions for counts, the Bernoulli distribution for presence–absence data, and the Tweedie and log-normal distributions for modelling non-negative continuous data such as biomass (Foster & Bravington 2013). Ordinal data are also permitted, with cumulative probit regression used.
- Construction of one or two dimensional plots of the latent variables via the `lvsplot` function. These are used for both model-based unconstrained ordination, including biplots, and residual ordination.
- A `get.enviro.cor` function for calculating correlations between species due to similarities in environmental responses, and a `get.residual.cor` function for calculating residual correlations accounted for by the latent variables, for example due to species competition and facilitation.
- Residual analysis for checking the validity of model assumptions, for example correct mean structure, distributional assumption and information criteria for selecting significant covariates, the number of latent variables and so on.

Pure latent variable models

We consider the well-known data set of counts of p=12 hunting spiders collected at n=28 sites in the Netherlands (Van der Aart & Smeenk-Enserink 1974). The data set is available in the R package MVABUND (Wang *et al.* 2012). As a first step to analysis, suppose we want to use unconstrained ordination to visualize the data set on a low-dimensional plot, looking for variation in species composition between sites as well as the indicator species characterizing the sites. The BORAL package implements a model-based approach to unconstrained ordination, by fitting a pure latent variable model.

$$\log(\mu_{ij}) = \alpha_i + \theta_{0j} + z_{i1} \times \theta_{j1} + z_{i2} \times \theta_{j2} = \alpha_i + \theta_{0j} + \boldsymbol{z}_i^T \boldsymbol{\theta}_j, \qquad \text{eqn 1}$$

where μ_{ij} is the mean response at site i for species j, θ_{0j} is the species-specific intercept, $\boldsymbol{z}_i = (z_{i1}, z_{i2})^T$ is a vector of two latent variables, and $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2})^T$ are the corresponding species-specific coefficients. We have included two latent variables so as to be able to construct a scatterplot of the ordinations; this is consistent with distance-based techniques like Non-metric Multidimensional Scaling (Kruskal 1964), where two ordination axes are typically chosen for data visualization. We have also included a row effect, α_i, to adjust for differences in site total abundance (Hui et al. 2014). This row effect is optional, but since our aim is to construct an ordination in terms of species composition rather than absolute abundance, then we have chosen to include it. In the BORAL package, the site effect may be included as either a fixed or normally distributed random effect, and we choose the former given n is small in this data set.

Unlike a generalized linear model where the covariates are observed, the latent variables $\boldsymbol{z}_i$ in equation (1) are unknown and therefore assumed to be random, drawn from a bivariate, standard normal distribution. When fitting the model, they are estimated simultaneously with the coefficients $\boldsymbol{\theta}_j$ and row effects α_i.

Assuming Poisson counts for the species, we can fit the above pure latent variable model in the BORAL package as follows,

```
library(mvabund); data(spider)
y <- spider$abun
fit.lvmp <- boral(y = y, family = "poisson",
num.lv = 2, row.eff = "fixed")
```

where `num.lv = 2` sets the number of latent variables to two, and `row.eff = "fixed"` inserts a fixed row effect. When the BORAL fitting function is run, a model file containing the JAGS script is created in the current working directory, which is then passed into JAGS for running MCMC estimation. As default prior distributions, BORAL assigns independent normal distributions with mean zero and variance 100 for θ_{0j}, uniform distributions between 0 and 50 for all non-negative parameters, for example overdispersion parameters in the negative binomial distributions as we shall use later, and normal distributions with mean zero and variance 20 for the latent variable coefficients $\boldsymbol{\theta}_j$ and α_i. Using a smaller variance for the latter tends to stabilize MCMC sampling, preventing it from sampling excessively large coefficients. This is especially helpful for small data sets such as the spider counts here; see Gelman *et al.* (2008) for related work on how weakly informative priors, even stronger than those imposed in the BORAL package, can generally be used to improve performance.

After fitting, a summary of the model can be obtained through `summary(fit.lvmp)`. Among other outputs, this returns posterior median estimates of the species-specific intercepts and latent variables (see Appendix S1 part 1). We also obtain 95% highest posterior density (credible) intervals for these parameters, accessed through `fit.lvmp$hpdintervals`. These may be interpreted as intervals containing the true parameter with 95% probability.

One of the advantages of a model-based approach to ordination is that we can straightforwardly perform residual analysis to assess the assumptions made in constructing the model; residual analysis and checking assumptions for distance-based methods is generally harder to perform, if possible at all. In the BORAL package, we can perform residual analysis using `plot(fit.lvmp)`, which produces a set of four graphs (Fig. 1). Overall, the plots indicated that there was overdispersion in the counts not captured by the Poisson distribution. To overcome this problem, we can use the negative binomial distribution instead, with quadratic mean–variance relationship variance$(y_{ij}) = \mu_{ij} + \phi_j\mu_{ij}^2$, where ϕ_j is the species-specific dispersion parameter, by setting `family = "negative.binomial"` in the above code. Residual analysis of the resulting fit, denoted here as `fit.lvmnb`, confirmed that the negative binomial distribution was a more appropriate fit to the spider counts (see Appendix S1 part 2).

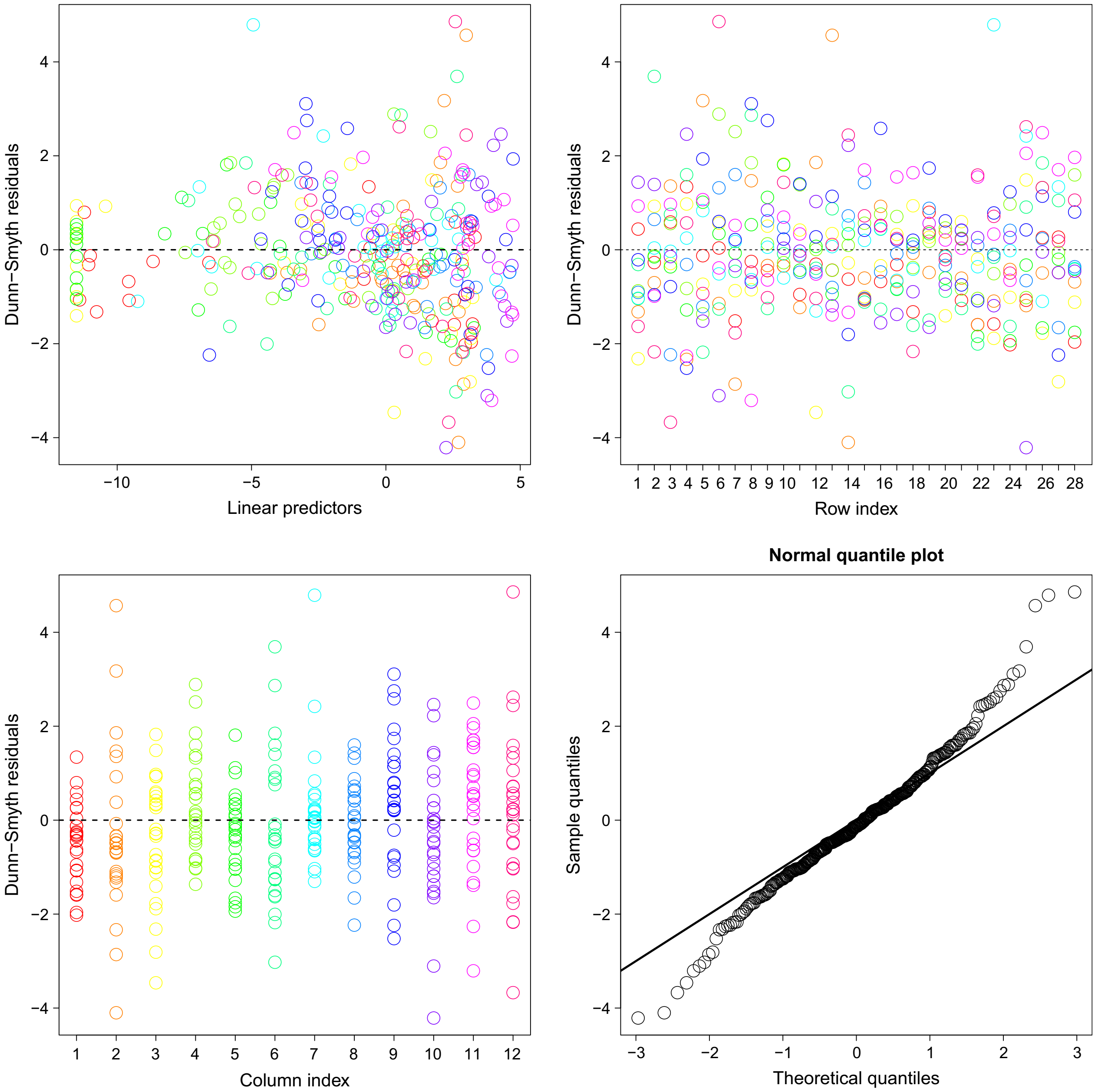

Fig. 1. Plots for residual analysis of the spider data set, returned from PLOT(FIT.LVMP). Each colour represents a different species. Top left: Dunn–Smyth residuals vs. linear predictors. Top right: Dunn–Smyth residuals vs. row index; Bottom left: Dunn–Smyth residuals vs. column index; Bottom right: normal quantile plot of Dunn–Smyth residuals. The presence of the funnelling effect in the top left plot indicates overdispersion in the data, with the normal quantile plot also suggesting that assuming Poisson counts is inappropriate.

Based on the pure latent variable model assuming negative binomial counts, we can construct an ordination plot using `lvsplot(fit.lvmnb)`. This produces a biplot based on the posterior median estimates (Fig. 2 left panel). By default, all species are shown on the biplot, although there is an `ind.spp` argument that can be used to plot only the most important or indicator species, as judged by the Euclidean norm of their coefficients $\boldsymbol{\theta}_j$. For the spider data set, the biplot informally distinguished three clusters of sites, with site 25 located close to the centroid of the three clusters. Sites 22–24, 26–28 were characterized by *Arctosa perita* (*Arctperi*) and to a lesser extent *Alopecosa fabrilis* (*Alopfabr*), while *Pardosa lugubris* (*Pardlugu*) strongly identified with sites 8, 15–21. Overall, the site pattern observed in the left panel in Fig. 2 matches those seen in Hui *et al.* (2015), which also compared ordinations between model-based and distanced-based methods. We point out, however, that Hui *et al.* (2015) focused only on ordinations of the sites, while the BORAL package offers biplots.

Correlated response models

To further the analysis of the spider counts, we study how well the assemblage as a whole (both species abundance and composition) is explained by the environment. This is done by fitting separate species generalized linear models with the covariates available in the data set. However, it is expected that species exhibit correlation beyond that due to possible similarities or differences in their environmental response, for example due to biotic interactions, and it is imperative that we take this into account to ensure our inferences remain valid in the presence of this residual correlation (Warton *et al.* 2015).

In the BORAL package, we can account for residual correlation by fitting a correlated response model, where latent variables are included alongside the measured covariates,

$$\begin{aligned}\log(\mu_{ij}) &= \theta_{0j} + \text{soil.dry}_i \times \beta_{j1} + \text{bare.sand}_i \times \beta_{j2}\\ &\quad + \text{fallen.leaves}_i \times \beta_{j3} + \text{moss}_i \times \beta_{j4}\\ &\quad + \text{herb.layer}_i \times \beta_{j5} + \text{reflection}_i \times \beta_{j6} + \boldsymbol{z}_i^T\boldsymbol{\theta}_j\\ &= \theta_{0j} + \boldsymbol{x}_i^T\boldsymbol{\beta}_j + \boldsymbol{z}_i^T\boldsymbol{\theta}_j,\end{aligned} \qquad \text{eqn 2}$$

where $\boldsymbol{x}_i$ is a vector denoting the six covariates (soil dry mass, per cent cover bare sand, per cent cover fallen leaves, per cent cover moss, per cent cover herb layer, reflection of the soil surface), and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{j6})^T$. Note that we have chosen to remove the row effect, as we are now using the environmental covariates directly to also explain the differences in site total abundance.

Rather than a method of model-based unconstrained ordination as in (1), the latent variables in equation (2) can be interpreted as a device to account any residual covariation not explained by the $\boldsymbol{x}_i$'s. This is similar to the residual covariance matrix included in the joint species distribution model of Pollock *et al.* (2014), but with the critical difference that the latent variables offer a more parsimonious method of modelling residual covariation. Specifically, Pollock et al. (2014) uses a random intercept u_{ij} in place of $\boldsymbol{z}_i^T\boldsymbol{\theta}_j$, where the vector $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ip})^T$ is drawn from a multivariate normal distribution with unstructured covariance matrix. Estimating all $p(p+1)/2$ elements of this covariance matrix becomes problematic in cases where p is large. By contrast, the use of two latent

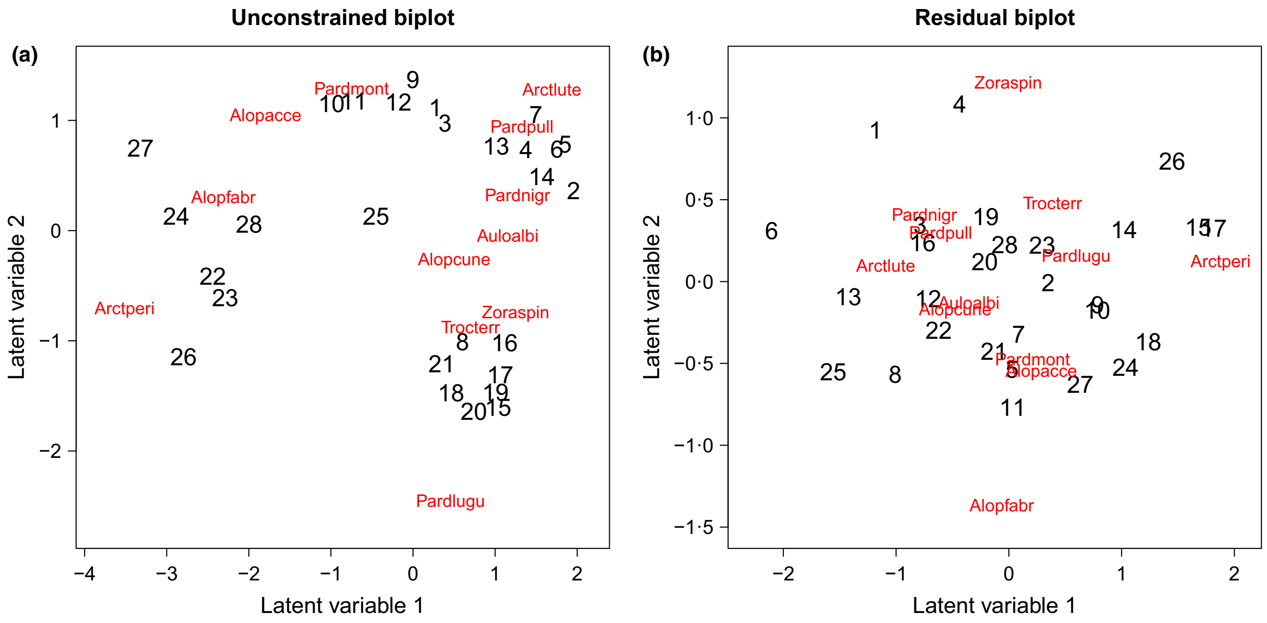

Fig. 2. Model-based unconstrained (left) and residual ordination (right) biplots for the spider data set, based on the posterior median estimates. Each of the 28 sites is labelled by its row number, while the 12 species are shown in red and labelled by their abbreviated names. For the unconstrained ordination, three clusters of sites can be discerned, with site 25 approximately located approximately in the middle of the three clusters. No pattern was observed in the residual ordination.

variables (statistically speaking, the residual covariance matrix is of rank 2) and $z_i^T\theta_j$ reduces the number of parameters down to approximately 2*p*.

One attractive way of thinking about the latent variables in the correlated response model is as missing predictors (Warton *et al.* 2015): we may have covariates that were not recorded and/or included in the model, and if these covariates are informative for the species response, then they will induce a residual correlation between the species. The inclusion of the latent variables serves directly to account for this, with the z_i's being the predicted values of the missing predictors and the θ_j's the corresponding coefficients. Indeed, if a low-rank residual covariance matrix is found to be appropriate, it suggests that the residual correlation between species is perhaps due to shared responses to unmeasured environmental predictors. Furthermore, from the predicted values of z_i, it might also be possible to hypothesize what these unmeasured or unanticipated environmental covariates are, which may coincide with additional, prior knowledge regarding the sites that was not included in the model (Harris 2015).

The correlated response model in equation (2) can be fitted in the BORAL package as follows,

```
X <- scale(spider$x)
fit.Xnb <- boral(y = y, X = X, family = "negative.
binomial", num.lv = 2,  save.model = TRUE)
```

The above is similar to the syntax for the pure latent variable model, except for the inclusion of a model matrix `X` and the row effect being omitted. Note also the use of `save.model = TRUE` to save the MCMC samples after fitting, which can then be accessed after fitting via `fit.Xnb$jags.model`; to conserve memory, the BORAL package does not save the MCMC samples by default. After fitting, we can again study the point estimates through `summary(fit.Xnb)`, use residual analysis through `plot(fit.Xnb)` to check model assumptions and obtain credible intervals through `fit.Xnb$hpdintervals` (see Appendix S1 part 3). In particular, the credible intervals obtained from the correlated response model lead to valid inference compared to if we were to assume no residual correlation, that is `num.lv = 0`, as they include additional uncertainty into the model due to residual covariation between species (a result confirmed in the simulations of Warton *et al.* 2015).

With the correlated response model, we are able to separate the correlations between species due to environmental response, $x_i^T\beta_j$, and the residual correlation due to other sources of covariation, $z_i^T\theta_j$. In the BORAL package, estimates of these correlations can be obtained and plotted as follows.

```
envcors <- get.enviro.cor(fit.Xnb)
rescors <- get.residual.cor(fit.Xnb)
library(corrplot)
corrplot(envcors$sig.cor, type = "lower",
diag = FALSE,
title = "Correlations due to covariates",
mar = c(3,0.5,2,1), tl.srt = 45)
corrplot(rescors$sig.cor, type = "lower",
diag = FALSE,
title = "Residual correlations",
mar = c(3,0.5,2,1), tl.srt = 45)
```

We have chosen to plot only the significant correlations, as based on the 95% credible intervals excluding zero. From the resulting plots (Fig. 3), we observed similar numbers of positive and negative correlations due to

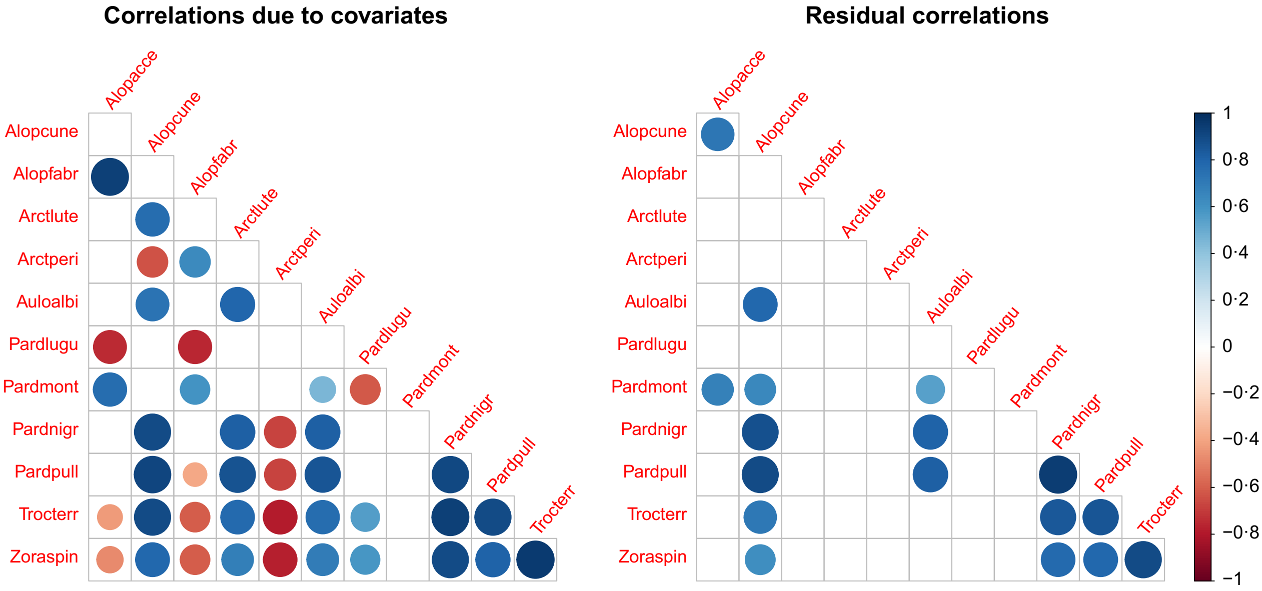

Fig. 3. Plots of the correlations between species due to the environmental response (left) and residual correlations (right), based on the correlated response model. Only the significant correlations, as based on the 95% credible intervals excluding zero, have been plotted. The sign of the correlations are represented by colours (red and blue for negative and positive correlations respectively), while the strength of correlations are represented by the size of the circles.

environmental response, while the residual correlation was dominated by a small number of strong, positive correlations. Another way to visualize the residual correlation between species is through a 'residual ordination' plot, and this can be constructed in the boral package using `lvsplot(fit.Xnb)` analogous to Section 1. The resulting biplot, displayed in the right panel of Fig. 2, did not exhibit any clustering pattern in the sites, while most species were grouped towards the middle of the plot.

Finally, one approach to quantify how much of the species co-occurrence is explained by covariates (that is, how well the predictor variables describe the species assemblage) is through differences in the trace of the estimated residual covariance matrix induced by the latent variables (Warton *et al.* 2015). From the above code, this can be obtained as `rescors$trace`. For the spider data set, when we compared a pure latent variable model (similar to the equation 1 but without site effects) to the correlated response model, the trace decreased from 178·92 to 107·92. This implies that environmental covariates accounted for approximately 40% of the covariation between species.

Other features

Aside from those seen above, there are other functionalities that the boral package offers. We conclude by discussing two of these.

INCLUSION OF TRAITS

If data on species traits are (also) available, these can be incorporated into correlated response models to explain differences in species environmental responses. Suppose for each species, we observe a vector of traits $\boldsymbol{t}_j$. The boral package can incorporate this by treating the coefficients θ_{0j} and $\boldsymbol{\beta}_j$ as random effects, drawn from a normal distribution with mean depending on the traits,

$$\theta_{0j} \sim N(\boldsymbol{t}_j^T \boldsymbol{\kappa}_0, \sigma_0^2) \quad \text{and} \quad \beta_{jk} \sim N(\boldsymbol{t}_j^T \boldsymbol{\kappa}_k, \sigma_k^2),$$

where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2, and β_{jk} is the k^{th} regression coefficient in $\boldsymbol{\beta}_j$. For k=0,1,. . ., the coefficients $\boldsymbol{\kappa}_k$ relate the species traits to the intercepts and regression coefficients, while σ_k^2 accounts for any extraneous variation not explained by traits. This random effects approach to incorporating traits into the model is similar to the mixed model approach proposed by Jamil *et al.* (2013), although residual covariation was not considered in that paper, and differs from the fixed effects methods in Brown *et al.* (2014), which could be regarded as a special case of the above when $\sigma_0^2 = \sigma_1^2 = \ldots = 0$.

In the boral fitting function, there is a `traits` argument that accepts a matrix of species traits, and a `which.traits` argument that determines which traits should be included into each of the means of the random effects normal distributions. An example illustrating its use is given in Appendix S2.

MODEL SELECTION

The boral package offers several information criteria for selecting the error distribution, important environmental covariates and so on, and these are available by applying the `summary` function. We caution, however, that the application of information criteria in hierarchical Bayesian models remains an active area of research (Gelman, Hwang & Vehtari 2014). As an alternative, the boral package provides a method known as Stochastic Search Variable Selection (George & McCulloch 1996) via the `ssvs.index` argument. Briefly, this works by augmenting the covariates with indicator variables, so that in each sample of the MCMC chain, the covariate may or may not be included depending on whether the indicator variable equals one or zero, respectively. After fitting, boral returns the posterior probability of the covariate being included in the model. For instance, results in Appendix S1 part 4 show that, based on the posterior probabilities from applying stochastic search variable selection to the spider data set, all six covariates are important for at least one species, with reflection of the soil surface and cover of fallen leaves being overall the most important covariates for explaining species responses.

Discussion

As research into statistical methodology for analysing multivariate abundance data progresses at a rapid pace, there is a strong need for software that makes these new methods available to ecologists. This article presents boral an r package that implements one form of hierarchical Bayesian modelling for multivariate abundance data. Key innovations of the boral package include its ability to handle a variety of response types, tools for constructing model-based unconstrained and residual ordinations, and the use of latent variables to parsimoniously model residual correlation between species.

In recent years, many papers have emerged advocating various types of model-based methods for community-level inference, with many of these papers having their own set of code or package (see Box 4 in Warton *et al.* 2015, for a table listing some of these methods). An important future work will therefore be to gather and compare the various software that has been produced, for example how is the boral package similar/different from other packages such as hmsc and mistnet? How do they compare in terms of performing community-level inference and computational burden in data sets where n and p are large? Another avenue of research would be to perform a comprehensive empirical study quantifying what the benefits and potential drawbacks of model-based approaches are compared to distance-based methods. Of course, the boral package itself is a work in progress, and in future versions, we hope to include a wider choice of (especially weakly informative) prior distributions, better methods for choosing the number of latent variables and automated tools for diagnosing convergence of MCMC chains.

Acknowledgements

This research was supported by the Australian Research Council discovery project grant DP140101259. Thanks to Sara Taskinen, Scott Foster and the UNSW Eco-Stats team for useful discussions.

Data accessibility

Both the hunting spider count data set used in this main text and the data set used in Appendix S2 are publicly available from the R package MVABUND (Wang *et al.* 2012).

References

Blanchet, F.G. (2014) *HMSC: Hierarchical Modelling of Species Community. R Package Version 0.6-2/r47*. URL http://rpackages.ianhowson.com/rforge/HMSC/.

Brown, A.M., Warton, D.I., Andrew, N.R., Binns, M., Cassis, G., & Gibb, H. (2014) The fourth-corner solution – using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution*, **5**, 344–352.

Clark, J. & Gelfand, A. (2006) *Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Applications: Statistical Methods and Applications*. Oxford University Press Inc., New York, USA.

Dunstan, P.K., Foster, S.D., Hui, F.K., & Warton, D.I. (2013) Finite mixture of regression modelling for high-dimensional count and biomass data in Ecology. *Journal of Agricultural, Biological and Environmental Statistics*, **18**, 357–375.

Foster, S.D. & Bravington, M.V. (2013) A Poisson–Gamma model for analysis of ecological non-negative continuous data. *Environmental and Ecological Statistics*, **20**, 533–552.

Gelman, A., Jakulin, A., Pittau, M.G., & Su, Y. (2008) A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, **2**, 1360–1383.

Gelman, A., Hwang, J., & Vehtari, A. (2014) Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, **24**, 997–1016.

George, E.I. & McCulloch, R.E. (1996) Stochastic search variable selection. *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson & D.J. piegelhalter), pp. 203–214. Chapman & Hall/CRC, Florida, USA.

Golding, N. & Harris, D.J. (2015) *BayesComm: Bayesian Community Ecology Analysis. R Package Version 0.1-2*.

Halstead, B.J., Wylie, G.D., Coates, P.S., Valcarcel, P., & Casazza, M.L. (2012) Exciting statistics: the rapid development and promising future of hierarchical models for population ecology. *Animal Conservation*, **15**, 133–135.

Harris, D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, **6**, 465–473.

Hui, F.K., Taskinen, S., Pledger, S., Foster, S.D., & Warton, D.I. (2015) Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, **6**, 399–411.

Jamil, T., Ozinga, W.A., Kleyer, M., & Ter Braak, C. (2013) Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *Journal of Vegetation Science*, **24**, 988–1000.

Kruskal, J.B. (1964) Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115–129.

Lunn, D.J., Thomas, A., Best, N., & Spiegelhalter, D. (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, **10**, 325–337.

Plummer, M. et al. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) March*, pp. 20–22. Vienna, Austria.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., Vesk, P.A., & McCarthy, M.A. (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, **5**, 397–406.

Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Elsevier Ltd., San Diego, USA.

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.

Walker, S.C. (2015) Indirect gradient analysis by Markov-chain Monte Carlo. *Plant Ecology*, **216**, 697–708.

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., Blanchet, F.G., O'Hara, R.O.O., Taskinen, S., Walker, S.C., & Hui, F.K.C. (2015) So many variables: joint modeling in community ecology. *Trends in Ecology and Evolution*, **30**, 766–779.

Received 25 September 2015; accepted 13 November 2015
Handling Editor: Timothée Poisot

Supporting Information

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

Appendix S1. Code and output relating to analysis of the spider data set.

Appendix S2. Code and output for using the BORAL package to fit a correlated response model, where species traits are included to help try and explain differences in species environmental response.