Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution doi: 10.1111/j.2041-210X.2012.00190.x

mvabund – an R package for model-based analysis of multivariate abundance data

Yi Wang[1,2], Ulrike Naumann[1], Stephen T. Wright[1], and David I. Warton[1,3]*

[1]*School of Mathematics and Statistics, The University of New South Wales, Sydney, NSW 2052, Australia;* [2]*School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW 2052, Australia; and* [3]*Evolution & Ecology Research Centre, The University of New South Wales, Sydney, NSW 2052, Australia*

Summary

1. The mvabund package for R provides tools for model-based analysis of multivariate abundance data in ecology.

2. This includes methods for visualising data, fitting predictive models, checking model assumptions, as well as testing hypotheses about the community–environment association.

3. This paper briefly introduces the package and demonstrates its functionality by example.

Key-words: community composition, generalised linear model, graphical methods, negative binomial regression, permutation test, resampling methods, significance test

In ecology, multivariate abundance data are widely used to study how community structure changes along environmental gradients and to test hypotheses about the impact of some environmental variable or experimental treatment. For example, Warwick, Clarke & Gee (1990) used a spatially blocked design to show that a disturbance treatment significantly affected a meiobenthic community consisting of 12 copepod species. However, the distance-based methods of analysis used by Warwick, Clarke & Gee (1990), still commonly used today, were unable to: (i) test whether the treatment effect was consistent across blocks or whether it operated differently in different blocks (test for interaction); (ii) identify which species expressed the treatment effect (multiple testing); (iii) predict the abundance of each species in different treatments or blocks (prediction). More recent distance-based methods (Clarke & Gorley 2006; Leathwick *et al.* 2011, for example) aim to address some of these issues, but they inherit from the distance-based framework some potentially serious problems in interpretability and performance (Warton, Wright & Wang 2012). This short paper introduces a new R package, mvabund, containing new methods of analysis that directly address all three issues listed above using a *model-based* framework.

There has been a recent trend towards model-based approaches to the analysis of multivariate abundance data in ecology (Yee 2010; Ives & Helmus 2011; Ovaskainen & Soininen 2011). The mvabund package builds on this trend by developing a novel set of hypothesis testing tools using the generalised linear models (GLM) framework (Warton 2011). This is a flexible and powerful framework for analysing abundance data – capable of handling most common data types (presence/absence, presence-only, count, etc.) and shown to have better power properties than distance-based methods (Warton 2011; Warton, Wright & Wang 2012). The key model fitting function is manyglm, which fits a separate GLM to each species, using a common set of explanatory variables. The anova and summary functions, which work on manyglm objects in the same way as for glm, use resampling-based hypothesis testing to make community-level and taxon-specific inferences about which factors or environmental variables are associated with the multivariate abundances. These inference tools take into account correlation between species, which is not possible using standard glm tools.

The main features of the mvabund package are new methods for visualising data, fitting predictive models, checking model assumptions, and testing hypotheses about the community-environment association. These features are summarised below.

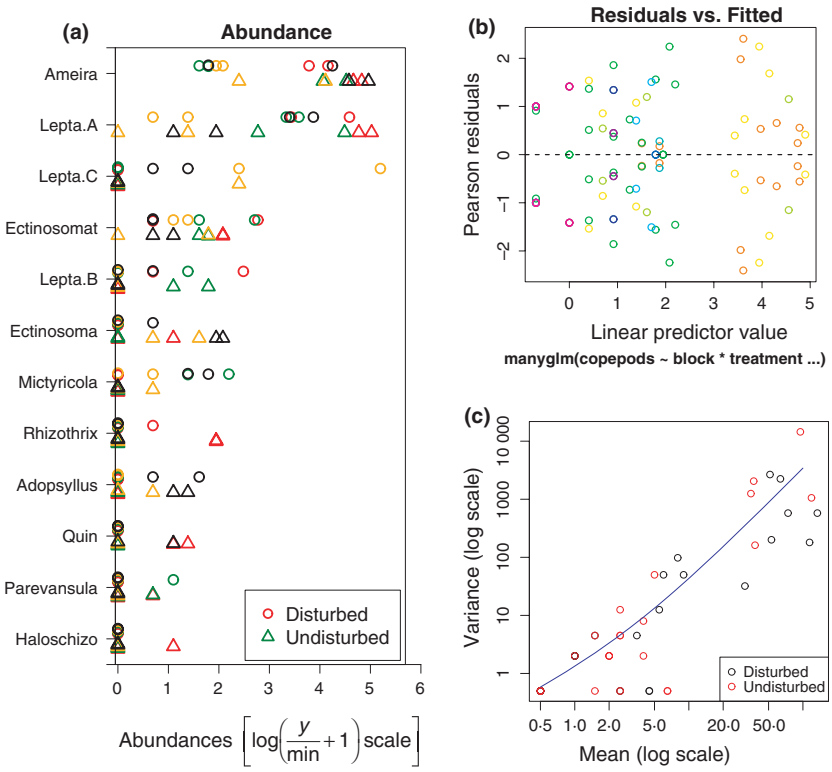
VISUALISING DATA

The package can produce a range of plots to visualise multivariate abundance data, as in Warton (2008). The following commands can be used to visualise the effect of treatment on copepod abundance (Fig. 1a), for the Tasmania data set:

```
> data(Tasmania)
> attach(Tasmania)
> tasmvabund < mvabund(copepods)
> plot(tasmvabund ~ treatment, col = as.
numeric(block))
```

There is a suggestion that treatment reduced abundance of *Ameira* and *Ectinosoma*, whereas it may have increased abundance of *Mictyricola*. We can test the hypothesis of a treatment effect, but first, we need to find a suitable model for the copepod data.

*Correspondence author. E-mail: david.warton@unsw.edu.au



(d) Analysis of Deviance Table

```
Model:  manyglm(copepods ~ block * treatment,
family = "negative binomial")

Multivariate test:
                  Res.Df   Df.diff   Dev   Pr(>Dev)
blocks                12         3 326.1      0.001 ***
treatment             11         1 106.5      0.008 **
blocks:treatment       8         3  48.5      0.063 .

Univariate Tests:
                  Ameira              Adopsyllus          Ectinosoma      ...   Mictyricola     ...
                  Dev    Pr(>Dev)     Dev    Pr(>Dev)     Dev    Pr(>Dev) ...   Dev    Pr(>Dev) ...
blocks            18.066 0.037        23.793 0.018        23.113 0.019    ...   19.15  0.033    ...
treatment         29.98  0.045         0.091 0.986        21.072 0.073    ...   21.072 0.071    ...
blocks:treatment  7.581  0.493         1.294 0.946         0.833 0.946    ...    5.269 0.493    ...
```

Fig. 1. (a) `plot.mvabund` plots copepod abundance against treatment groups, samples in the same block have been coded with the same colour. (b) `plot.manyglm` produces a residual vs. fits plot to check the quadratic mean–variance assumption of negative binomial regression (with different species coded in different colours); little pattern suggests the assumption is plausible. (c) `meanvar.plot` produces a mean–variance plot; there is a roughly quadratic trend. (d) `anova.manyglm` produces an analysis of deviance table. The multivariate test indicates a significant treatment effect and non-significant interaction. Separate results for each species (adjusted for multiple testing) are also returned.

FITTING PREDICTIVE MODELS

Predictive models can be fitted using the `manyglm` function:

```
> tas.nb < manyglm(copepods ~ block* treatment,
family = "negative.binomial")
```

This function fits a generalised linear model (Zuur, Ieno & Elphick 2010) separately to each species. The argument `family` specifies the assumed distribution of the data. Negative binomial regression was specified in the above (`family = "negative.binomial"`), this often being appropriate for count data, with the mean–variance function tending to be quadratic rather than linear (O'Hara & Kotze 2010). Other available options include `binomial` (for presence–absence data), `Poisson` (for presence-only data) and `gaussian`. The formula `copepods`

~ block*treatment specifies an orthogonal two-factor model. That is, the model for the number of copepods of species j found at site i (Y_{ij}) is negative binomial:

$$Y_{ij} \sim NB(\mu_{jkl}, \phi_j) \qquad \text{eqn 1}$$

where site i is in block k and it received treatment l. The overdispersion parameter ϕ_j is constant across sites but can vary across species, and the mean of Y_{ij} is μ_{jkl}, a log-linear function of block and treatment:

$$\log(\mu_{jkl}) = \text{intercept}_j + \text{block}_{jk} + \text{treatment}_{jl} + \text{block} \times \text{treatment}_{jkl} \qquad \text{eqn 2}$$

An important feature of using a model-based analysis framework is that the model that is fitted can be used for predictive purposes (Ives & Helmus 2011). For example, the predicted values for each species and site can be obtained using:

```
> predict(tas.nb, type = "response")
```

Other key commands familiar to R users for exploring glm objects are also available for manyglm objects, e.g. coef(tas.nb), residuals(tas.nb).

A key model assumption is independence: the Y_{ij} are assumed to be independent (conditionally on block and treatment) across sites, and there is also an implicit assumption of independence across species in separately applying maximum likelihood estimation to each species. This latter assumption is relaxed in hypothesis testing, as described later.

Independence of sites, as always, is an important assumption in multivariate analysis which can only be ensured through an appropriate study design (Gotelli & Ellison 2004). The remaining model assumptions , however, can be checked from the data as below.

CHECKING MODEL ASSUMPTIONS

There are two key assumptions in any manyglm fit: the mean–variance assumption, specified by choice of the family argument as in eqn 1; and the assumed relationship between mean abundance and environmental variables, as specified by the link function and formula as in eqn 2. The appropriateness of these assumptions can be checked by plotting the residuals vs. fits (Fig. 1b):

```
> plot(tas.nb)
```

Little pattern suggests that the quadratic mean–variance assumption implicit in using family = "negative binomial" is plausible.

A second way to study the mean–variance relationship is to plot it directly:

```
> meanvar.plot(copepods~tr.block, col = as.
numeric(treatment))
```

where tr.block is a factor variable containing the eight block×treatment combinations. This function plots the sample variance against the sample mean for each species within each factor level. A quadratic line (as in Fig. 1c) fits this mean–variance trend well.

The second assumption, the log-linearity assumption of eqn 2, was unimportant for the copepod example because the model included orthogonal factors only. However, if quantitative variables are included in the model (e.g. pH), then a trend in size of residuals at different fitted values (e.g. a 'U-shape' in Fig. 1b) would suggest a violation of the log-linearity assumption.

TESTING HYPOTHESES ABOUT THE COMMUNITY–ENVIRONMENT ASSOCIATION

Multivariate hypotheses about the treatment effect and treatment-by-block interaction can be tested:

```
> anova(tas.nb, p.uni = "adjusted")
```

which returns a table testing the significance of each term in the log-linear model of eqn 2 (Fig. 1d). It can be seen that there is a significant effect of the treatment factor (Dev = 106·5, $P = 0{\cdot}008$), meaning that treatment has a significant multiplicative effect on mean abundance. The interaction between blocks and treatments is not significant (Dev = 48·5, $P = 0{\cdot}063$), meaning that the multiplicative treatment effect is consistent across blocks. The p.uni argument allows univariate 'species-by-species' results to be returned as well, some of which are displayed in Fig. 1c. These P-values have been adjusted to control the family wise error rate across species, using a resampling-based implementation of Holm's step-down multiple testing procedure (Westfall & Young 1993). It can be seen that *Ameira*, *Ectinosoma* and *Mictyricola* have large treatment effect test statistics (above 20), consistent with the pattern seen in Fig. 1a, but their adjusted P-values are only marginally significant ($P_{adj} =$ 0·045, 0·071, 0·073, respectively). This demonstrates a key advantage of multivariate analysis – there is greater power to detect patterns when analysing all species simultaneously ($P = 0{\cdot}008$) than when looking for a pattern separately in each species ($P_{adj} \geq 0{\cdot}045$).

The argument test specifies the test statistic used, which can be "LR" (likelihood ratio), "wald" or "score" (Warton 2011). By default, the multivariate test statistic is calculated assuming independence of species response variables, which makes the test statistic a simple sum of the univariate test statistics across species, as in Warton *et al.* (2012). This assumption can be relaxed using the cor.type argument to use statistics that account for correlation between variables, which improves power properties of the test statistic (Warton 2011) but at the expense of computation time. The statistics in mvabund are analogous in construction to conventional MANOVA and multivariate regression statistics, indeed setting family = "normal" and cor.type = "R" results exactly in conventional multivariate statistics. While other choices of family allow for non-normal data, other choices of cor.type produce statistics that better handle situations where there are many response variables compared to the number of sites, as for the Tasmania data set (Warwick, Clarke & Gee 1990). Note, however, that irrespective of choice of cor.type, inferences are valid even when abundances are correlated across taxa – because the significance of the test statistic is evaluated via resampling rows of data, which preserves and accounts for any correlation structure across species within sites.

The argument `nBoot` sets the number of resamples used to estimate the *P*-value (default is 1000), and `resamp` decides which resampling method is used to calculate the *P*-values. Available methods are case, residual and score resampling, residual permutation or parametric bootstrap (`resamp = "case", "resid", "score", "perm.resid", "montecarlo"`) (Davison & Hinkley 1997). We have studied all of these methods by simulation and found them to provide approximately valid tests for typical count data sets, while we specifically recommend the parametric bootstrap for presence/absence data. All the resampling methods are computationally intensive, so the core functions are coded in C++ with an R/C++ interface (Eddelbuettel & Francois 2011) to reduce computation time. Most analyses are completed in seconds or minutes, but longer computations times would be expected for data sets with hundreds or thousands of sites and/or species. In such cases, it is advisable to first estimate computation time using an initial run with a small number of bootstrap samples, for example, `nBoot = 50`.

The `anova` function is quite flexible and handles nested hypotheses in the standard way:

```
> tas.block < manyglm(copepods ~ block,
family = "negative.binomial")
> anova(tas.block, tas.nb)
```

tests whether the block×treatment model (`tas.nb`) explains any additional variation not captured by a block model alone.

The `mvabund` package is available on CRAN (`cran.r-project.org`) and is compatible with version 2.13 of R and above.

Acknowledgements

This research was supported under Australian Research Council's Discovery Projects funding scheme (project number DP0987729). Thanks to Warwick, Clarke & Gee (1990) for making their data available and to the handling editor and anonymous reviewers for suggestions that improved the manuscript and R package.

References

Clarke, K.R. & Gorley, R.N. (2006) *PRIMER v6: User Manual/Tutorial*. PRIMER-E, Plymouth.

Davison, A.C. & Hinkley, D.V. (1997) *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

Eddelbuettel, D. & Francois, R. (2011). Rcpp – seamless R and C++ integration. *Journal of Statistical Software*, **40**, 1–18.

Gotelli, N.J. & Ellison, A.M. (2004) *A Primer of Ecological Statistics*. Sinauer Associates, Sunderland, Massachusetts.

Ives, A.R. & Helmus, M.R. (2011) Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs*, **81**, 511–525.

Leathwick, J.R., Snelder, T., Chadderton, W.L., Elith, J., Julian, K. & Ferrier, S. (2011) Use of generalised dissimilarity modelling to improve the biological discrimination of river and stream classifications. *Freshwater Biology*, **56**, 21–38.

O'Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in Ecology & Evolution*, **1**, 118–122.

Ovaskainen, O. & Soininen, J. (2011) Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, **92**, 289–295.

Warton, D.I. (2008) Raw data graphing: an informative but under-utilized tool for the analysis of multivariate abundances. *Austral Ecology*, **33**, 290–300.

Warton, D.I. (2011) Regularized sandwich estimators for analysis of high dimensional data using generalised estimating equations. *Biometrics*, **67**, 116–123

Warton, D.I., Wright, S.T, & Wang, Y. (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, **3**, 89–101.

Warwick, R.M., Clarke, K.R. & Gee, J.M. (1990) The effect of disturbance by soldier crabs Mictyris platycheles H. Milne Edwards on meiobenthic community structure. *Journal of Experimental Marine Biology and Ecology*, **135**, 19–33.

Westfall, P. & Young, S. (1993) *Resampling-Based Multiple Testing*. John Wiley & Sons, New York, New York.

Yee, T.W. (2010) The VGAM package for categorical data analysis. *Journal of Statistical Software*, **32**, 1–34.

Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology & Evolution*, **1**, 3–14.

Received 26 October 2011; accepted 16 January 2012
Handling Editor: Robert Freckleton