Received: 25 September 2018 | Revised: 13 December 2018 | Accepted: 8 January 2019

DOI: 10.1002/ece3.4948

ORIGINAL RESEARCH

WILEY Ecology and Evolution Open Access

A pathway for multivariate analysis of ecological communities using copulas

Marti J. Anderson[1,2] | Perry de Valpine[3] | Andrew Punnett[2] | Arden E. Miller[4]

[1]New Zealand Institute for Advanced Study (NZIAS), Massey University, Auckland, New Zealand

[2]PRIMER-e (Quest Research Limited), Auckland, New Zealand

[3]Department of Environmental Science, Policy and Management, University of California, Berkeley, California

[4]Department of Statistics, University of Auckland, Auckland, New Zealand

Correspondence
Marti J. Anderson, New Zealand Institute for Advanced Study (NZIAS), Massey University, Auckland, New Zealand.
Email: m.j.anderson@massey.ac.nz

Funding information
Royal Society of New Zealand

Abstract

We describe a new pathway for multivariate analysis of data consisting of counts of species abundances that includes two key components: copulas, to provide a flexible joint model of individual species, and dissimilarity-based methods, to integrate information across species and provide a holistic view of the community. Individual species are characterized using suitable (marginal) statistical distributions, with the mean, the degree of over-dispersion, and/or zero-inflation being allowed to vary among a priori groups of sampling units. Associations among species are then modeled using copulas, which allow any pair of disparate types of variables to be coupled through their cumulative distribution function, while maintaining entirely the separate individual marginal distributions appropriate for each species. A Gaussian copula smoothly captures changes in an index of association that excludes joint absences in the space of the original species variables. A permutation-based filter with exact family-wise error can optionally be used a priori to reduce the dimensionality of the copula estimation problem. We describe in detail a Monte Carlo expectation maximization algorithm for efficient estimation of the copula correlation matrix with discrete marginal distributions (counts). The resulting fully parameterized copula models can be used to simulate realistic ecological community data under fully specified null or alternative hypotheses. Distributions of community centroids derived from simulated data can then be visualized in ordinations of ecologically meaningful dissimilarity spaces. Multinomial mixtures of data drawn from copula models also yield smooth power curves in dissimilarity-based settings. Our proposed analysis pathway provides new opportunities to combine model-based approaches with dissimilarity-based methods to enhance understanding of ecological systems. We demonstrate implementation of the pathway through an ecological example, where associations among fish species were found to increase after the establishment of a marine reserve.

KEYWORDS

abundance data, discrete counts, over-dispersion, species' associations, statistical model, zero-inflation

1 | INTRODUCTION

Multivariate ecological community data, consisting of counts of species' abundances, have a number of salient statistical properties that have been studied over many decades (Bliss & Fisher, 1953; ter Braak, 1996; Clarke, Chapman, Somerfield, & Needham, 2006; Martin et al., 2005; McArdle, Gaston, & Lawton, 1990; Taylor, Woiwood, & Perry, 1979; Whittaker, 1952). They are typically high dimensional (Dunstan, Foster, Hui, & Warton, 2013)—the number of species (p) often exceeds the number of sampling units (N), and most species are rare (McGill et al., 2007), contributing many zeros to an ($N \times p$) matrix of count data (**Y**). In addition, individual variables generally show over-dispersion (McArdle et al., 1990; White & Bennetts, 1996) and often zero-inflation (Martin et al., 2005; Welsh, Cunningham, Donnelly, & Lindenmayer, 1996; Wenger & Freeman, 2008). The degree of over-dispersion and zero-inflation varies, not only among species (Clarke, Chapman et al., 2006; Taylor, 1961), but also within a species across different environmental conditions, temporally or spatially (McArdle & Anderson, 2004; Smith, Anderson, & Millar, 2012; Taylor et al., 1979). Thus, multi-species datasets consist of mixed variable types; counts of different species are generally incommensurable, due to differences in the sizes, morphologies, life-history strategies, detectabilities, and behaviors of different species.

Species also display associations with one another (Somerfield & Clarke, 2013). Statistical non-independence may reflect phylogenetic or functional inter-relationships (Paradis & Claude, 2002), synchronous (or asynchronous) behavior or dispersal mechanisms (Kendall, Bjørnstad, Bascompte, Keitt, & Fagan, 2000), or inter-specific interactions, such as competition (Goldberg & Landa, 1991), commensalism, or trophic relationships (Zurell, Pollock, & Thuiller, 2018). Associations can also be generated indirectly through species responding in similar (or opposing) ways to environmental gradients or habitats (Dunstan, Foster, & Darnell, 2011; Warton et al., 2015). Furthermore, relationships among species vary in time and space under changing biotic or abiotic conditions (Clark, Wells, & Lindberg, 2018); for example, species may only compete when resources are limiting (Perry, Mitchell, Zutter, Glover, & Gjerstad, 1994), in certain parts of their range (Pacala & Roughgarden, 1985), or in the absence of predation (Chase et al., 2002).

To analyze multi-species count data, many researchers have used methods based on an $N \times N$ matrix of dissimilarities, **D**, among sampling units (Anderson, 2001; Clarke, 1993). These approaches handle high-dimensional count data and reliably detect important changes in the structure of ecological communities (Anderson, 2001; Clarke, 1993; Clarke, Somerfield, & Chapman, 2006; Legendre & De Cáceres, 2013). The focus here is to measure holistic changes in the identities of species and potentially also changes in species' relative or proportional abundances. Dissimilarities, often calculated using measures that exclude joint absences, such as Bray–Curtis, Hellinger, or Jaccard (Legendre & Legendre, 2012), integrate information across all species to define ecological relationships among sampling units. Some measures (such as Gower's measure, 1971) also accommodate data having different types of variables (see Legendre & Legendre, 2012). Most dissimilarity measures of interest to ecologists emphasize the extent to which two sampling units either share, or do not share, any species in common. Thus, important community-level concepts such as beta diversity (Anderson et al., 2011; Vellend, 2001) and turnover (Baselga, 2010), for example, are measured using dissimilarities. However, dissimilarity-based methods create no formal model of the original variables. Roles of individual species and relationships among them are not directly identifiable, as no species-specific parameters are estimated. Hence, one cannot predict the makeup of communities under defined scenarios, nor readily calculate power.

To characterize ecological communities and make species-level predictions under specified hypotheses, formal joint statistical models of the species variables are required. The multi-faceted challenge for developing such models is to deal simultaneously with (generally) over-dispersed, zero-inflated, high-dimensional, inter-related, mixed sets of (usually discrete) variables, including a host of rare species, for which no single multivariate statistical distribution can be readily articulated. A successful model of community data should allow a wide variety of species-specific marginal distributions that can flexibly change in time and space, while accounting for meaningful inter-specific associations.

There has been rapid recent development of new statistical models for multivariate species data that also incorporate inter-specific associations (Clark, Nemergut, Seyednasrollah, Turner, & Zhang, 2017; Clark et al., 2018; Golding & Purse, 2016; Harris, 2015, 2016; Hui, 2016; Hui, Taskinen, Pledger, Foster, & Warton, 2015; Nieto-Lugilde, Maguire, Blois, Williams, & Fitzpatrick, 2017; Niku, Warton, Hui, & Taskinen, 2017; Pollock et al., 2014; Popovic, Hui, & Warton, 2018; Thorson et al., 2016; Warton et al., 2015). Several of these, such as stochastic feed-forward neural networks ("*mistnet*"; Harris, 2015), Bayesian Gaussian process models (GP SDMs; Golding & Purse, 2016), or Markov random fields (MRF; Clark et al., 2018), have so far been used only on presence–absence data to enhance species distribution models (SDMs, Elith & Leathwick, 2009), although extensions to abundance data may well be feasible.

Models of multivariate abundance data include generalized linear models (GLMs) with (typically, for counts) a log link and either a Poisson or negative binomial (NB) error, and with correlations between species modeled using generalized estimating equations (GEEs, Wang, Naumann, Wright, & Warton, 2012; Warton, 2011; Warton & Guttorp, 2011). Alternatively, relationships among species can be modeled parsimoniously by including latent random variables as linear predictors in the GLM—called generalized linear latent variable models (GLLVMs, Hui et al., 2015; Niku et al., 2017; Warton et al., 2015). The latent variables are intended to capture correlations due to unmeasured environmental drivers or biotic interactions, and one can use model selection to identify an appropriate number of latent variables to include in the model (Hui et al., 2015). More sophisticated GLLVMs can also accommodate spatial/temporal autocorrelation (Ovaskainen, Roy, Fox, & Anderson, 2016; Thorson et al., 2016, 2015), the inclusion of species' traits/phylogenies (Ovaskainen

et al., 2017), or variation in correlations at different hierarchical spatial scales (Ovaskainen, Abrego, Halme, & Dunson, 2016).

Generalized joint attribute models (GJAMs) have also been proposed for modeling multivariate ecological data (Clark et al., 2017). GJAMs model covariances between mixed variable types (presence–absence, ordinal, discrete, or continuous) on their original scales. Discrete data (such as counts) are modeled via censoring, with partitions and weights chosen to allow linkages between different variable types (Clark et al., 2017). Partition widths and associated effort in interval censoring also can be chosen arbitrarily to accommodate mean–variance relationships. Imputation across censored intervals maps discrete variables into a multivariate normal (MVN) space to estimate covariances, with Bayesian analysis being used to estimate the latent (imputed) states.

We consider that copulas (Mai & Scherer, 2017) also hold great promise for flexible joint modeling of multivariate ecological count data (Popovic et al., 2018). A copula is a function representing a joint distribution as a mapping from the cumulative distribution functions (cdf) of its marginals, hence can be used to couple virtually any pair of variables (Mai & Scherer, 2017; Sklar, 1959). Copulas allow a tailored multivariate distribution to be constructed from two separate parts: (a) the univariate marginal distributions for each variable; and (b) the joint distributions of the variables in a multivariate copula space. Although implemented fairly widely in other fields (Nikoloulopoulos & Karlis, 2009; Shi & Valdez, 2014), copulas have not yet been widely used in ecology (but see de Valpine, Scranton, Knape, Ram, & Mills, 2014; Popovic et al., 2018).

Our primary motivation for using copulas for ecological count data is that they allow any marginal distribution to be used for any variable. One does not need to forego the utility of the wide potential array of existing univariate statistical distributions, each with its own interpretable parameters, to build a joint model. Also, in copula models, associations among variables are modeled separately from their marginal distributions, making them easy to interpret, and the particular copula distribution used to model associations can also be flexibly chosen to fit a specific context. In contrast, latent variable models typically confound correlation structures with marginal distributions, as correlations among species are induced *via* latent variables that in turn alter the marginal distributions.

Here, we shall restrict our attention to Gaussian (MVN) copula distributions, and also to counts of species (abundance data) arising from one-way ANOVA-type designs, but the core ideas are readily extended to other types of mixed datasets, other copula distributions, and/or more complex sampling/experimental designs. Gaussian copulas can draw on the rich statistical literature surrounding MVN distributions, while tailoring marginal distributions to non-normal ecological variables. Fortunately, difficulties in estimating parameters for Gaussian copulas with discrete marginals (Faugeras, 2017; Genest & Nešlehová, 2007) have recently been surmounted (see Appendix 1).

The aim of this work is twofold. First, we provide an accessible description of copulas and show how they can work for ecological count data through a simple bivariate example. Second, we outline a pathway for the analysis of multivariate ecological count data that combines the use of copulas with dissimilarity-based methods. More specifically, copulas are first used to characterize the properties of individual variables and their associations in a formal parametric statistical model. Dissimilarity-based methods are then used to examine community-level patterns for whole assemblages of species that have been simulated from these copula models under defined scenarios.

Our analysis pathway (a) characterizes each individual species via estimation of marginal distributions and their associated parameters; (b) addresses high dimensionality (optionally) by screening data to identify significant pair-wise associations, using an index that excludes joint absences; (c) characterizes associations among species via estimation of a copula model and its associated parameters; and (d) proposes simulation from copula models to generate realistic ecological data under specified null or alternative hypotheses for model-based inference, ordination, and power analysis in dissimilarity-based settings. In the proposed pathway, we allowed both the marginal parameters and the copula parameters to vary across a priori groups, to maximize flexibility.

We demonstrate the analysis pathway with an example dataset: counts of fishes (p = 47 species) from the Poor Knights Islands, New Zealand (two of the 47 species are shown in Figure 1). Sampling occurred at three different times (September 1998: n_1 = 15, March 1999: n_2 = 21, and September 1999: n_3 = 20), spanning the establishment of a no-take marine reserve in October 1998 (Willis & Denny, 2000; data are provided in Supporting Information Table S1

FIGURE 1 Two fish species found at the Poor Knights Islands, New Zealand: *Chirodactylus spectabilis* (top) and *Parma alboscapularis* (bottom). Photographs by Paul Caiger

and are also available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.3s6rm0f).

2 | A GAUSSIAN COPULA FOR DISCRETE NON-NORMAL DATA

A copula is defined by Sklar's seminal theorem (1959). Let F be a p-dimensional distribution function with margins F_1, F_2, ..., F_p. There exists a p-dimensional copula C such that for all $(y_1, y_2, ..., y_p) \in \mathbb{R}_p$ the following holds:

$$F(y_1, y_2, \ldots, y_p) = C(F_1(y_1), F_2(y_2), \ldots, F_p(y_p)) \quad (1)$$

C is unique if $F_1, F_2, ..., F_p$ are continuous. Conversely, if C is a p-dimensional copula and F_1, F_2, ..., F_p are univariate distribution functions, then the function $F(y_1, y_2, ..., y_p)$ is a p-dimensional distribution function (Mai & Scherer, 2017).

To understand how copulas work, consider that the position of an individual value y of a random variable Y having probability density function (pdf) $f_Y(y)$ is able to be expressed as a value along its cdf, denoted $F_Y(y)$, on the interval [0,1]. This provides a direct mapping of values from one pdf to another via their cdfs. This approach can also be used to map a value drawn from the pdf of a continuous random variable to the probability mass function (pmf) of a discrete random variable (Figure 2a). For example, consider a random variable $Y \sim$ Poisson ($\mu = 2.5$) and a standard normal variable $Z \sim N(\mu = 0, \sigma^2 = 1)$, whose cdf we will denote by $C_Z(z)$. A random value z drawn from Z can be mapped on to Y uniquely by taking the inverse function: $y = F_Y^{-1}(C_Z(z))$. Thus, suppose we draw $z = 0.524$, then $C_Z(0.524) = 0.70$; that is, we have drawn the 70th percentile of

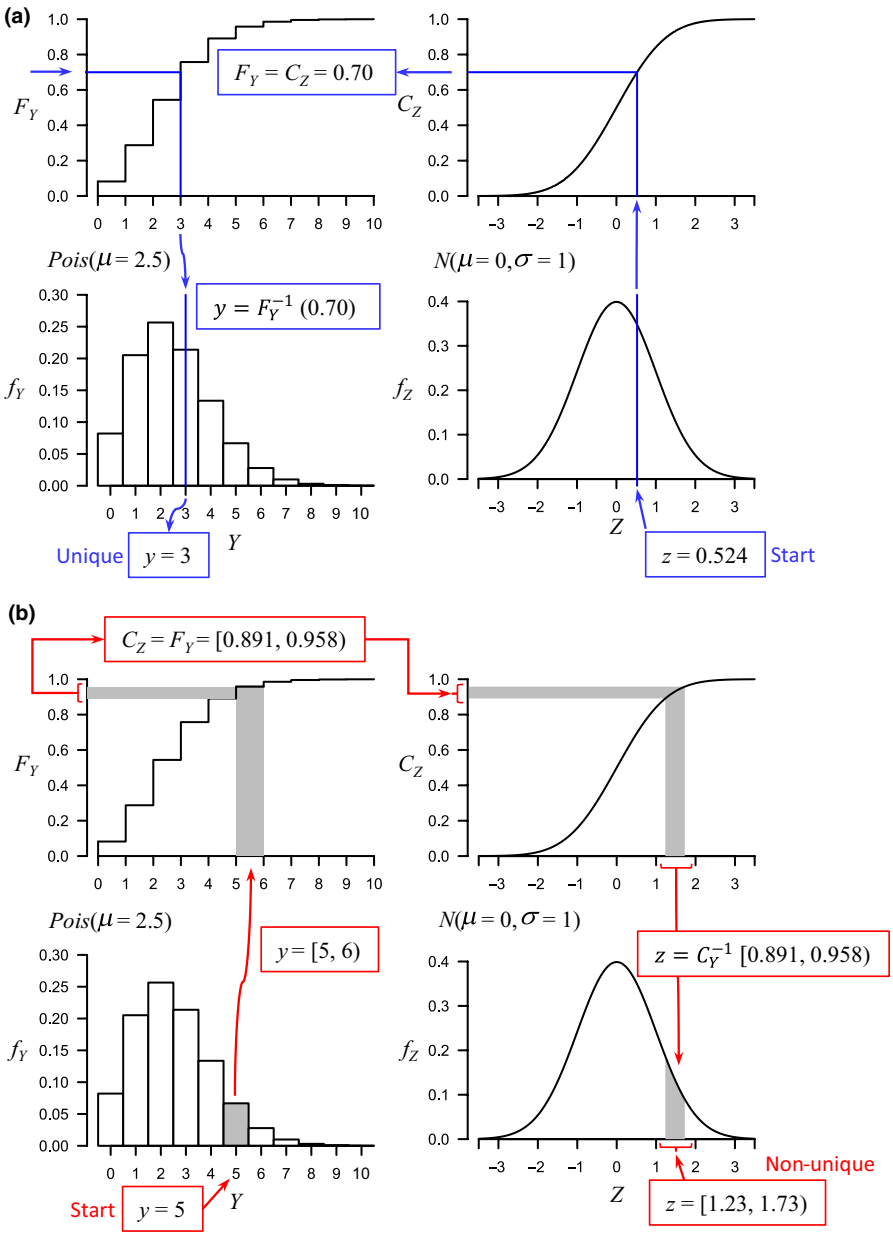


FIGURE 2 Schematic diagram showing the mapping between the probability density function of a continuous variable and the probability mass function of a discrete variable via their respective cumulative distribution functions (cdfs); (a) the mapping from the continuous to the discrete yields a unique value; (b) the mapping from the discrete to the continuous is non-unique, but produces a range of values in the space of the continuous variable

$f_Z(z)$. Mapping $F_Y(y)$ one-to-one on $C_Z(z)$, the 0.70 quantile of $f_Y(y)$ is then given by $y = F_Y^{-1}(0.70) = 3$ (Figure 2a, blue arrows).

This simple mapping idea is exploited for modeling the joint distribution (association) between any pair of disparate types of variables in a copula. Here, we shall use a Gaussian copula to handle multiple associated variables simultaneously (but see Mai and Scherer (2017); Shi and Valdez (2014) for a broader variety of copula distributions). Consider a bivariate ecological example, where Y_1 are counts of individual fish of the species *Chirodactylus spectabilis* (red moki), and Y_2 are counts of *Parma alboscapularis* (black angelfish) from the Poor Knights Islands (Figure 1). Model selection (using AICc) performed separately on each variable for the full set of data ($N = 56$) suggested the NB ($\mu = 2.714$, $\theta = 1.635$) and the zero-inflated NB (ZINB; $\mu = 15.375$, $\theta = 1.857$, $\pi = 0.095$) are suitable marginal distributions to model Y_1 and Y_2, respectively (Figure 3). (Note: we have simply posited here that the parameters for each of these marginal distributions are equivalent to their maximum likelihood estimates.)

A measure of association between two species that excludes joint absences is the index of association (Somerfield & Clarke, 2013; Whittaker, 1952):

$$I_{k\ell} = 1 - \frac{1}{2}\sum_{i=1}^{N}\left|\left(\frac{y_{ik}}{\sum_{j=1}^{N} y_{jk}}\right) - \left(\frac{y_{i\ell}}{\sum_{j=1}^{N} y_{j\ell}}\right)\right|$$

where y_{ik} denotes the count for species k in sampling unit i and $I_{k\ell}$ denotes the association between species k and species ℓ. For red moki and black angelfish, there is a statistically significant positive association of $I = 0.698$ ($p = 0.0001$, 10,000 permutations).

To model this association, we can use a standard bivariate normal distribution for the copula function (with variables Z_1 and Z_2) having correlation parameter $\rho = 0.574$ (lower left panel, Figure 3). A random sample of $\mathbf{z} = (1.28, 0.67)$ in the copula space (Figure 3, lower left) can be mapped into the bivariate space of species count data (Figure 3, upper right) by taking the inverse of the corresponding cdf values (the 90th and the 75th percentile, respectively) on each marginal distribution, that is, $\mathbf{y} = \left(F_{Y_1}^{-1}(0.90), F_{Y_2}^{-1}(0.75)\right)$, hence $\mathbf{y} = (30, 4)$. A large number of such random draws from the copula model will generate count data in the species space that preserves their association as well as their individual (and disparate) marginal distributions (Figure 3). For a given pair of variables, there is a smooth monotonic relationship between rho (ρ) in the copula space and the index of association (I) in the space of the original variables (Figure 4a), highlighting the utility of Gaussian copulas in ecological research. In contrast, Pearson correlations (r) calculated among the original variables do not show a strong relationship with the index of association (Figure 4b), as the former do not omit joint-absence information (Somerfield & Clarke, 2013).

One important complication, however, is the fact that a single point in the discrete data space corresponds to an entire region (a hyper-rectangle) in the copula space (Figure 2b). Thus, estimation of the copula parameter(s) from discrete datasets is problematic, as the mapping of discrete values into the copula space is non-unique. We can address this by performing Monte Carlo integration over each discrete interval (Shi & Valdez, 2014). Computational efficiency is achieved through a Monte Carlo expectation maximization (MCEM) algorithm (Wei & Tanner, 1990). Estimation of the parameters of the correlation matrix for a Gaussian copula with discrete marginal distributions is described in greater detail in Appendix 1.

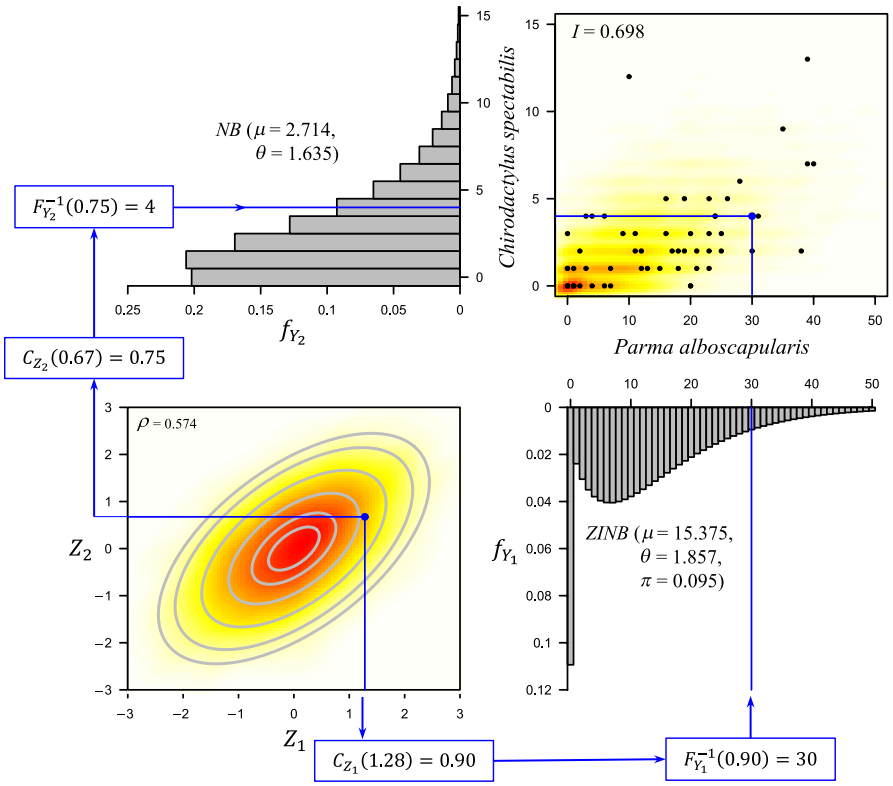


FIGURE 3 Standard bivariate normal (Gaussian) copula model with discrete marginal distributions for two fish species from the Poor Knights Islands. Points drawn from the copula space with correlation parameter $\rho = 0.574$ generate, through the specified negative binomial (NB) and zero-inflated NB marginal distributions, points in the discrete bivariate data space of counts that correspond to an expected inter-specific index of association of $I = 0.698$

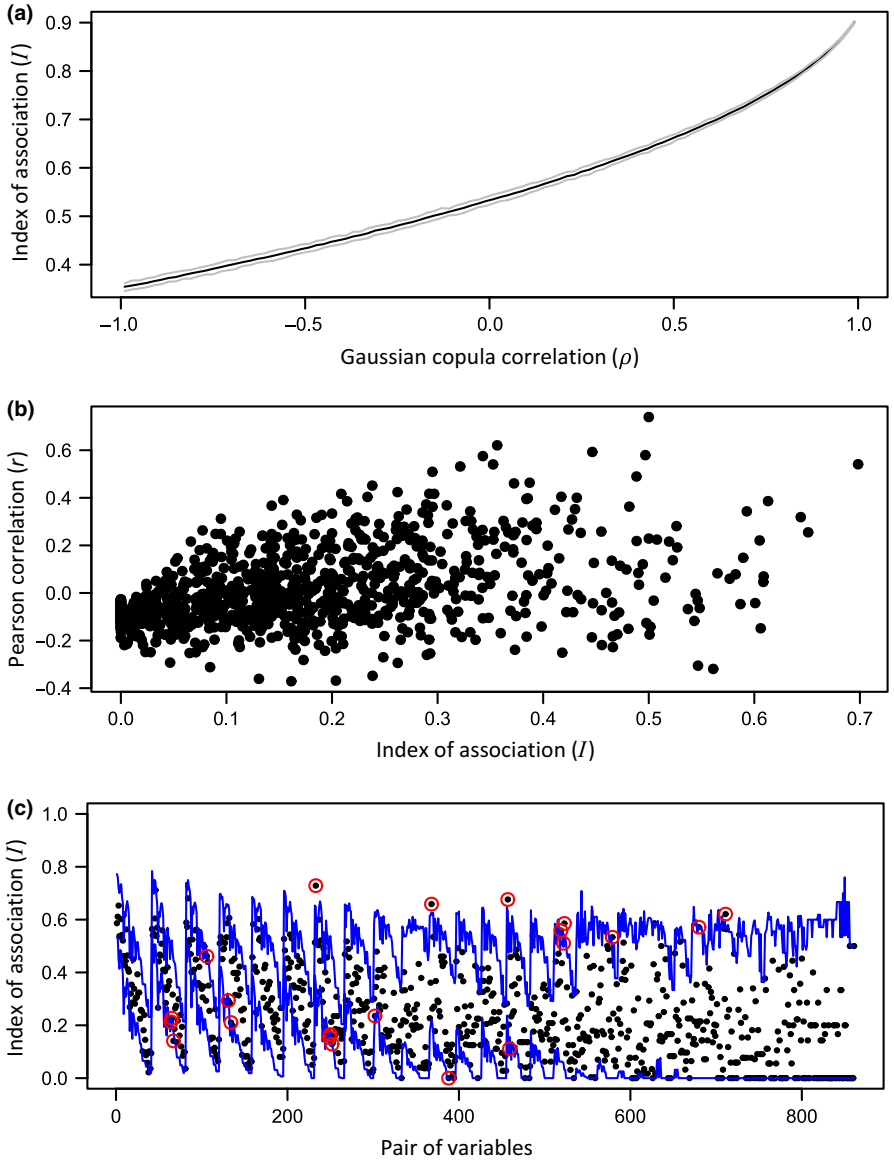


FIGURE 4 (a) Relationship between the index of association (I) in the discrete bivariate data space, with marginal distributions of negative binomial (NB) (μ = 2.714, θ = 1.635) and zero-inflated NB (μ = 15.375, θ = 1.857, π = 0.095; see Figure 3), as a smooth function of the correlation parameter (ρ) in the standard bivariate normal (Gaussian) copula space. The black line follows the mean and the gray lines follow the upper 0.975 and lower 0.025 quantiles of the distribution of 100 values of I that were each calculated using a sample of 5,000 simulated datasets from the multivariate normal copula distribution at each value of ρ (taken at 0.02-unit intervals between −1.0 and 1.0). (b) Relationship between the Pearson correlation (r) and the index of association (I) for all pairs of 47 variables (counts of fish species) from the Poor Knights Islands. (c) Index of association (I) between every pair of variables (black dots) for the fish data from the Poor Knights Islands for Time 2 only (p = 42, as five species did not occur at Time 2). Pairs are shown along the x-axis in decreasing order of species' importance, defined as frequency of occurrence. Blue lines show the upper and lower bounds from the permutation distribution of I, specific to each pair, for a two-tailed per-comparison empirical error rate of 0.01, obtained using 99,999 permutations. Red circles identify statistically significant associations

3 | A COPULA MODEL FOR ECOLOGICAL COUNT DATA

We propose the following steps to develop a full copula-based model for high-dimensional ecological count data:

i. *Identify appropriate marginal distributions*. For each species, an information criterion (such as AIC or AICc) may be used to choose among potential marginal distributions. Here, we restricted our attention to the following count distributions: Poisson, ZIP, NB, or ZINB (see Supporting Information Data S1 for R code). This step may include, for efficiency, identification of rare species that do not contain enough information to allow estimation of associations. Species may be flagged as "rare" if they occur as singletons or in only a small percentage of sampling units (e.g., <5%). We used AICc to identify marginal distributions for each of the p = 47 fish species in the Poor Knights dataset (Supporting Information Table S1), allowing both

the statistical distribution and estimated parameters to vary across the three groups.

ii. *Identify significant associations among species to model.* An index of association is calculated between every pair of species, the null hypothesis of no association is tested for each pair using *P*-values obtained by permutations (Somerfield & Clarke, 2013), and the significance level for the tests is suitably adjusted for multiple tests (see Supporting Information Data S2 for R code). For efficiency, species flagged as "rare" may (optionally) be omitted. For the adjustment, one may use, for example, an exact family-wise error rate (FWER), empirically derived from the full set of permutation distributions (Wheldon, Anderson, & Johnson, 2007), or a more conservative per-comparison error

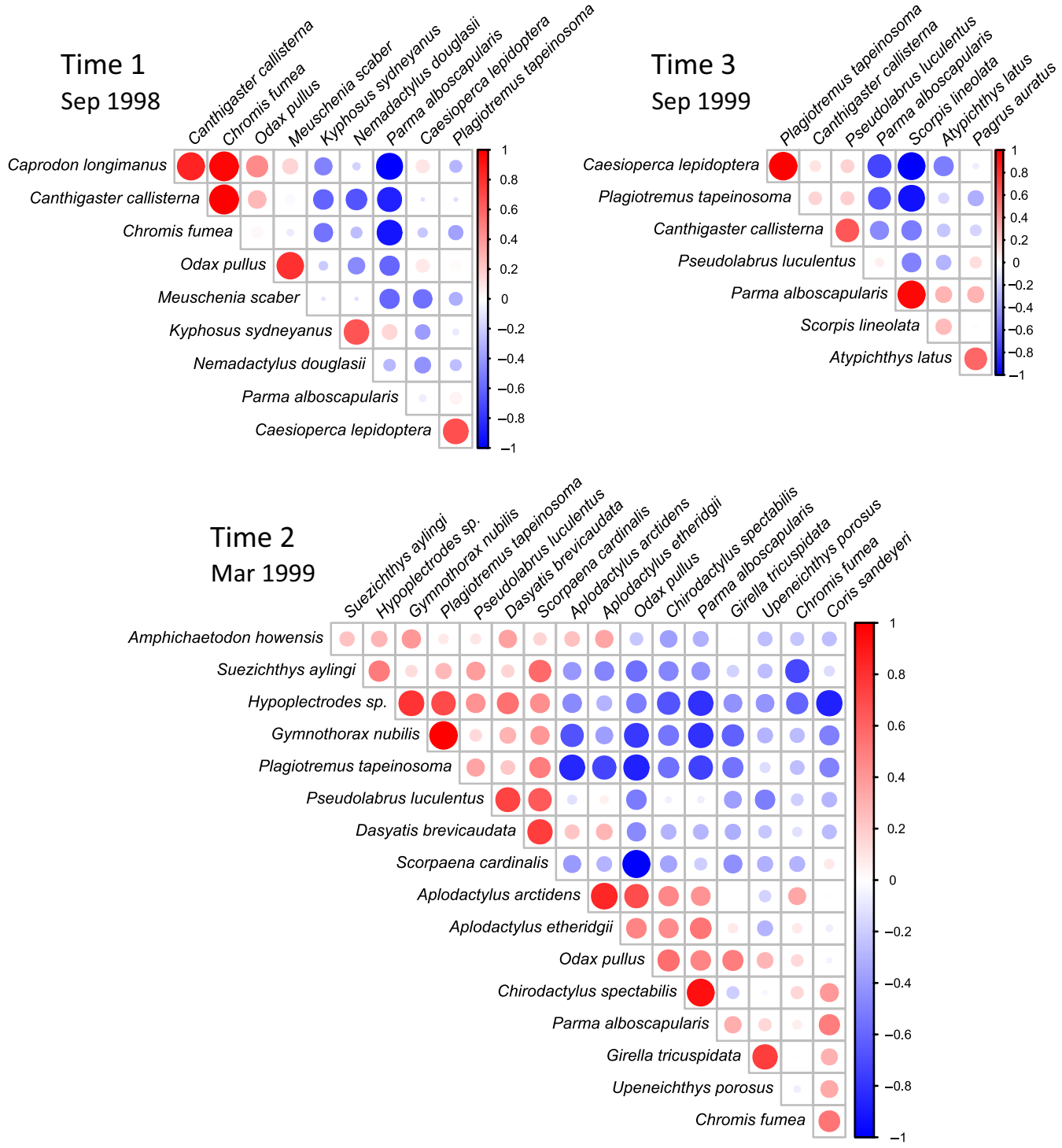


FIGURE 5 Heat maps of estimated copula correlation parameters for samples obtained at each of three different times (September 1998, March 1999, and September 1999) among fish species from the Poor Knights Islands that showed at least one statistically significant index of association (using a per-comparison error rate of 0.01 as a filter, e.g., see Figure 4c)

rate (PCER). We used a PCER of 0.01 for the fish data. This was done separately for each group. We identified five significant associations (involving 10 species) in September 1998, followed by a sharp increase to 20 associations (involving 17 species) in March 1999 (Figure 4c), and subsequent decrease to five associations (involving eight species) in September 1999. This step acts as a filter to reduce the size of the estimation problem for building a copula model and omits joint absences in the assessment of species' associations. It is, however, optional; one could allow the copula model in step (iii) to include all inter-specific associations.

iii. *Build a copula model and estimate its parameters*. We may reduce the problem to a subset of species, $m \le p$, showing a significant association with any other species. Given this subset of m species identified in step (ii), and each of their marginal distributions from step (i), estimate the parameters of the Gaussian copula's correlation matrix (see Appendix 1 for details; R code is provided in Supporting Information Data S3). We condition the estimation of copula correlation parameters on the fixed marginal distributions, which is both practical (ensuring marginal distributions fit each individual variable well) and efficient (Joe, 2005). Here, we estimated a separate copula correlation matrix for each group. Note that we may have chosen not to implement step (ii) above, or perhaps, even though step (ii) may reduce dimensionality dramatically ($m \ll p$), we may still have $N < m$, or m may begin to approach N such that some form of regularization is still desirable. In such cases, as we are using Gaussian copulas, a variety of methods for regularizing the inverse covariance matrix may be considered (Friedman, Hastie, & Tibshirani, 2008; Schäfer & Strimmer, 2005; Ullah & Jones, 2015; Yuan & Lin, 2007); for simplicity, we shall not pursue the topic of regularization further here. Furthermore, we hasten to add that neither rare species nor unassociated species are omitted from the copula models that follow, but they are presumed to be independent of other species. For the fish dataset, species' associations varied through time, and structured groups of associated species were easily seen in copula correlation matrices (Figure 5). There was an increase in the strength of associations after the establishment of the

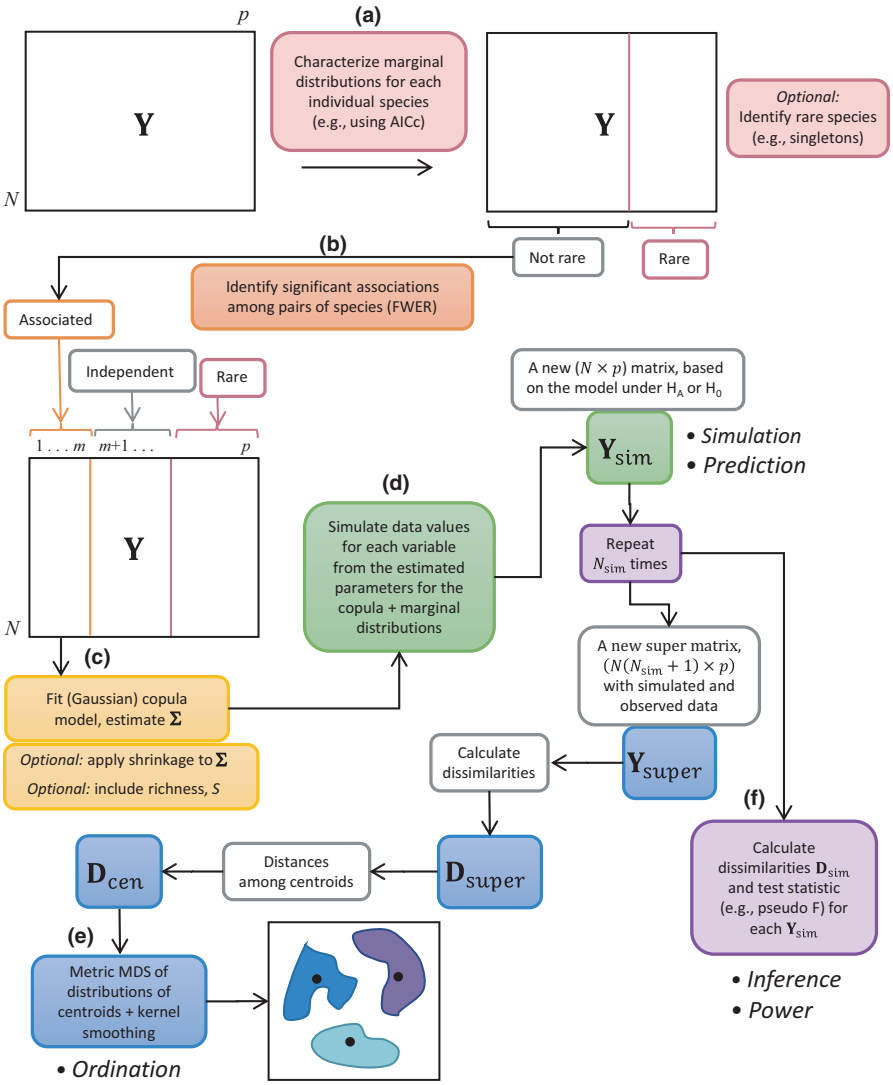


FIGURE 6 Schematic diagram of an overall pathway for analyzing ecological community data consisting of the following steps: (a) characterize marginal distributions for each species; (b) identify associations to model; (c) fit the copula; (d) simulate data under the full copula-based model for predictions under null or alternative hypotheses; (e) visualize dissimilarities among sampling units or centroids under the model using robust ordination techniques; and (f) calculate power for test statistics of interest

marine reserve (March 1999), which later subsided (September 1999). This is unlikely to have been a seasonal effect, as only three of the 10 species that showed significant associations in September 1998 did so in the following September 1999 (Figure 5).

4 | SIMULATE DATA AND VISUALIZE RESULTS

An overall pathway to model, simulate, and visualize ecological community data using copulas is shown schematically in Figure 6. Once the parameters for all of the marginal distributions and the correlation matrix of the copula have been estimated for a given group (Figure 6a–c), then one can readily draw a random sample of a given modeled community, $\mathbf{y}_{sim}$ (a vector of length p), for that group (Figure 6d) as follows:

1. Suppose there are $i = 1, \ldots, g$ groups. Let the subset number of species in group i showing a significant association with any other species in that group be denoted by m_i and their estimated $(m_i \times m_i)$ correlation matrix in the Gaussian copula be denoted by $\hat{\boldsymbol{\Sigma}}_{[m_i]}$. Expand this to obtain a $(p \times p)$ correlation matrix $\hat{\boldsymbol{\Sigma}}_i$ for group i by placing zeros in the remaining off-diagonal elements (corresponding to species that will be modeled as independent of one another), and 1s along the remaining diagonal elements.
2. Draw a random sample (a vector of length p) from the p-dimensional Gaussian copula distribution whose correlation matrix is $\hat{\boldsymbol{\Sigma}}_i$. Map the values obtained for each dimension in the copula space through the cdf of the individual marginal distribution for each species to obtain a simulated count value for each of the p species in the community.

Ordination plots of simulated data with original data can be used as a simple diagnostic to assess how sensible the model may be. What is usually of greater interest, however, is to consider changes in community structure among groups in the high-dimensional space articulated by the model. For this, we desire an ordination of the group centroids, along with some measure of the expected variation in those centroids (based on the model). This can be examined on the basis of any dissimilarity measure of interest.

Suppose there are n_i sampling units in group i and $N = \sum_{i=1}^{g} n_i$. One generates a new $(N \times p)$ matrix of simulated data $\mathbf{Y}_{sim}$ under the full copula model by drawing the n_i sampling units, consisting of p-length vectors $\mathbf{y}_{sim}$, for each group in accordance with that group's estimated copula correlation matrix $\hat{\boldsymbol{\Sigma}}_i$, marginal distributions, and associated parameters. This is then repeated N_{sim} times (where N_{sim} is typically somewhat large, say $N_{sim} = 100$). One can then construct a super-matrix $\mathbf{Y}_{super}$ of dimension $(N(N_{sim} + 1) \times p)$ which (row-wise) stacks the original matrix ($\mathbf{Y}$) together with all of the $\mathbf{Y}_{sim}$ matrices obtained *via* simulation under the model. From this, a chosen dissimilarity measure is calculated to yield $\mathbf{D}_{super}$. We wish to map the $N_{sim} \times g$ centroids for all of the groups from every simulated dataset along with the g original centroids onto an ordination diagram. We can do this by calculating the distances among the $(N_{sim} + 1) \times g$ centroids from the $\mathbf{D}_{super}$ matrix directly (Anderson, 2017) to obtain $\mathbf{D}_{cen}$. Metric multi-dimensional scaling (mMDS) can be used to visualize the distributions of centroids for each group under the model, along with the original group centroids. Kernel density contours (Duong, 2007) clarify the shapes of these distributions in the ordination space (Figure 6e).

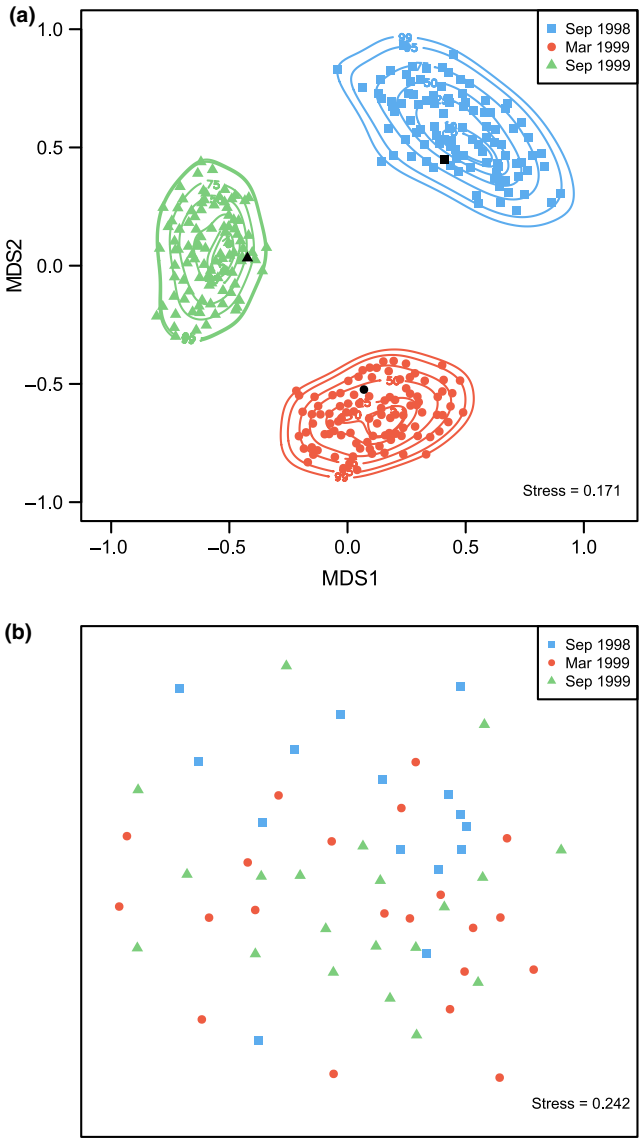


FIGURE 7 Ordinations based on Bray–Curtis dissimilarities of square-root-transformed abundances of fishes (47 species) from the Poor Knights Islands at three different times, obtained using: (a) metric multi-dimensional scaling (mMDS) of distances among centroids for the original data (black symbols), also showing centroids of 100 datasets (colored symbols) generated under the full copula model (parameters estimated separately for each group, shown as three different colors) along with kernel density contours; and (b) non-metric MDS plot of the original data, with replicate sites in each of three groups shown with three different colors

Using this approach for the Poor Knights' dataset, the observed centroid for each group falls well within the distribution of centroids for that group generated under the copula-based model (Figure 7a). The dispersion/shape of centroid distributions in the reduced-space ordination also clearly varied among groups. This graphic is far more informative than the (typically drawn) non-metric MDS plot of the original data (Figure 7b). The latter is dominated by large residual variation within each group and concomitant high stress, which masks group differences. Consider how, in the univariate analysis of data from ANOVA-type study designs, one typically plots the means for each group, along with their associated standard errors (which sensibly assumes normality for the distribution of means under the central limit theorem), to visualize both the relative positions and the variation in the group means. Similarly, the ordination in Figure 7a depicts the centroid for each group (in the space of the chosen resemblance measure) along with a visualization of the multivariate variation in the positions of those centroids (shown as density contours) under the assumptions of the full copula model. By "full copula model," we mean the set of estimated correlations among all pairs of species in the MVN copula space along with the full set of individual (in this case discrete) marginal distributions for each species and their associated estimated parameters.

It must be borne in mind, however, that Figure 7a was drawn under the assumption of a highly specific alternative hypothesis (H_A). The three groups have been asserted to be different, and all of the model parameters (copula plus marginals) have been estimated separately for each group. It is therefore not surprising that these three regions do not overlap with one another in the ordination space. We might also choose to visualize distributions of centroids under a true null hypothesis (H_0). For example, we can calculate centroids obtained under random permutation of the sampling units among the three groups. These permutation-based centroids assert the null hypothesis that sampling units are fully exchangeable among the groups to be true. We can examine the distributions of centroids under H_0 and also under H_A in a single ordination plot (Figure 8a). By including the originally observed centroids for each of the three groups here as well, we are able to gain an understanding of the position of our own data with respect to H_0 and the specific H_A that arises from these copula models (Figure 8a).

One might well ask: what would such a plot look like if the null hypothesis were true? Specifically, suppose we do the following: (a) take the full set of $N = 56$ sampling units and estimate a single set of marginal and copula parameters from these data (acknowledging no a priori groups, so H_0 is true); (b) generate three groups of "mock" data (with sample sizes of $n_1 = 15$, $n_2 = 21$, and $n_3 = 20$) directly from that model, then treat this dataset as if it were our "observed" data, but here we know that H_0 is actually true; (c) estimate marginal and copula parameters separately for each of these "groups" in our mock dataset; and then (d) simulate data and draw distributions of centroids under H_A and H_0 in the same way as was done for Figure 8a.

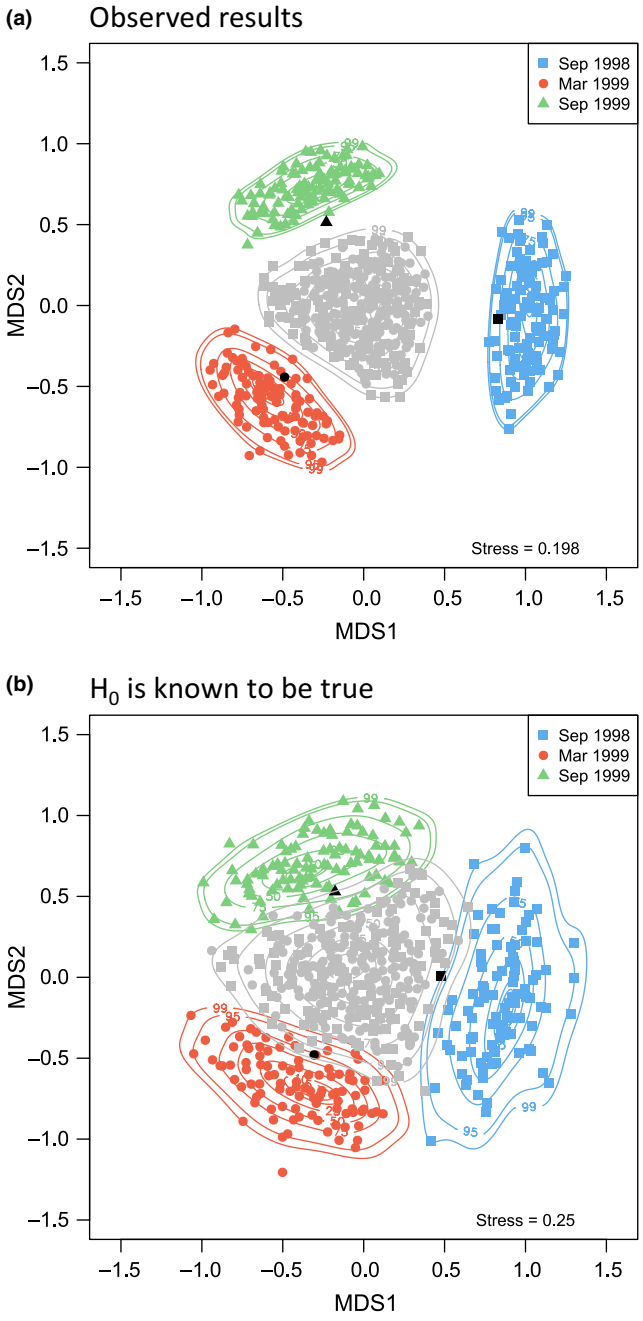


FIGURE 8 Ordinations based on Bray–Curtis dissimilarities of square-root-transformed abundances of fishes (47 species) from the Poor Knights Islands at three different times, obtained using: (a) metric multi-dimensional scaling (mMDS) of distances among centroids for the original data (black symbols) along with centroids of 100 datasets generated under three separate copula models for each of the three groups (colored symbols, H_A is true) and centroids of 100 datasets obtained by random permutation of the sampling units among the three groups (gray symbols, H_0 is true); (b) mMDS generated in the same manner as in (a), but where data consisted of "mock" observations where H_0 was known to be true (see text for details)

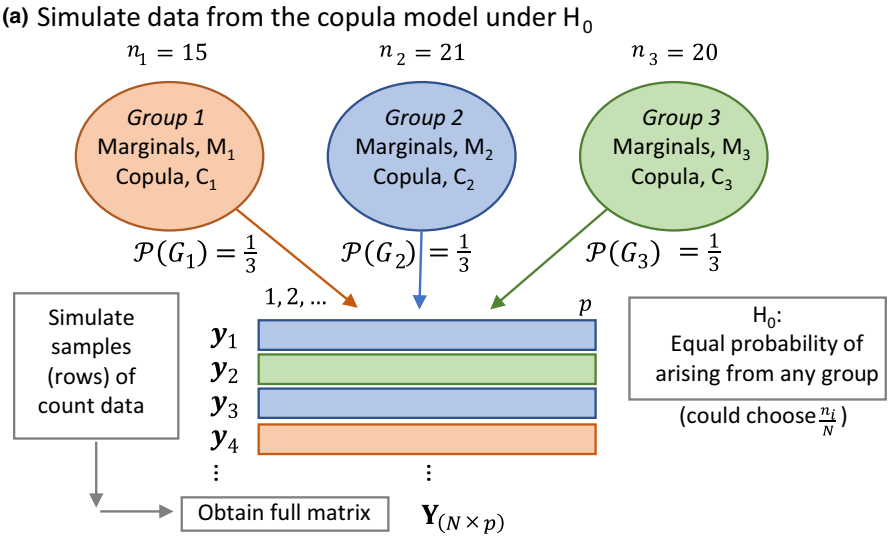
For the mock dataset (Figure 8b), the distributions of centroids for the three groups generated under H_A (three different colors) do appear separate from one another. This is because they arose

from three different sets of estimated parameters, even though the groups themselves, as we know, were, in this case, completely arbitrary. However, quite tellingly, the mock "observed" centroids (black symbols) also lie within the distribution of centroids under H_0 (gray symbols). The overlapping of contours for centroid distributions drawn under H_A with those drawn under H_0 also suggests a lack of real difference among the groups. Note too that there is quite high stress here (>0.24)—yet another signal that there is no distinctive group structure to display (Figure 8b). This can be contrasted with the clear group structure apparent in Figure 8a that was constructed based on the real data; we see no overlap in the centroid distributions under H_A with those under H_0, and the positions of the genuine observed centroids clearly favor H_A.

5 | MODEL-BASED INFERENCE AND POWER

5.1 | Model-based inference

We may generate data under a specified null hypothesis (H_0) to achieve model-based inference. One might consider a null hypothesis that asserts there are no groups and estimate a single set of parameters (marginals plus copula) for the full set of data. However, armed with a full copula model, having estimated separate parameters for each group, we may instead assume a simple null hypothesis that every sampling unit has an equal probability of arising from any group. Thus, we can generate $\mathbf{Y}_{sim}$ under H_0 such that each vector $\boldsymbol{y}_{sim}$ is drawn under a multinomial with probabilities of $1/g$ for each



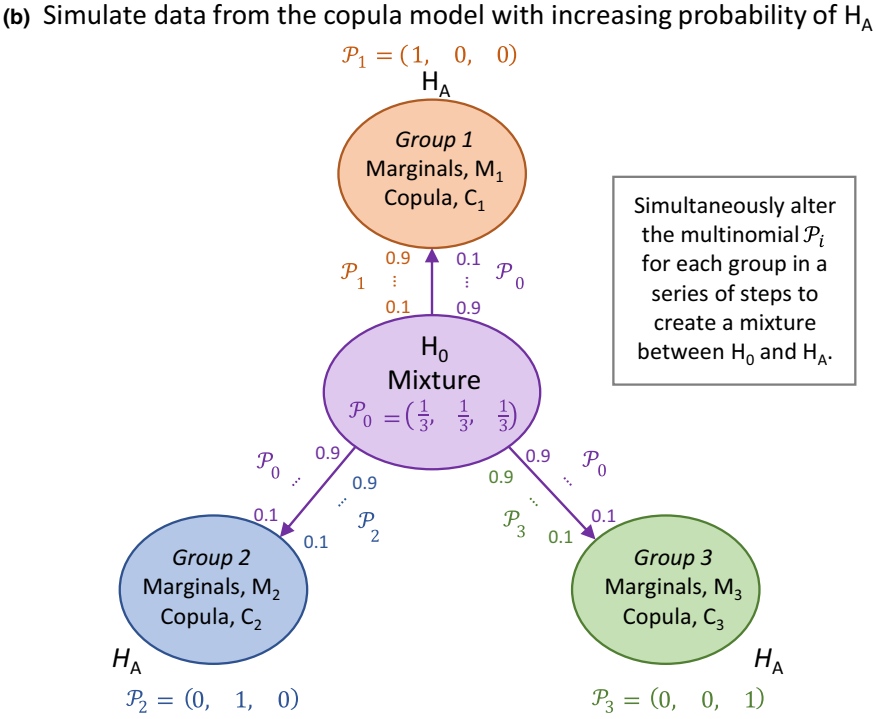


FIGURE 9 Schematic diagram showing methods of simulation using multinomial mixtures from copula models (with given parameters for the marginal, *M*, and copula, *C*, distributions) under (a) a null hypothesis of every sampling unit having an equal probability of arising from any of the groups; or (b) a specific alternative hypothesis for the case of three a priori groups of sampling units and where the alternative hypothesis asserts that all three groups are different from one another

group (Figure 9a). For the Poor Knights' data, these probabilities are $\mathscr{P}_i$= ⅓ for the i = 1, …, 3 groups. Another alternative would be to base null hypothesis probabilities on sample sizes: $\mathscr{P}_i = n_i/N$, appropriate, for example, if sample sizes directly reflected encounter rates of the existing groups in nature.

With $\mathbf{Y}_{sim}$ generated in this way, one may calculate a test statistic, such as the PERMANOVA pseudo-*F*, based on a chosen dissimilarity measure. Repeating this procedure many times generates a model-based distribution of pseudo-*F* under H_0. A *p*-value for model-based inference is calculated directly as the proportion of pseudo-*F* values under H_0 that equal or exceed the observed value. In the present example, the observed value of pseudo-*F* is 2.716 (vertical line in Figure 10a). A model-based *p*-value (from 4,999 random draws) is p = 0.0002. The assumptions here are that the specified copula and marginal distributions provide a realistic joint model for these data. The usual permutation-based test of pseudo-*F* is distribution-free, so is preferable for robust inference (in this case, with 4,999 permutations, it yielded an identical *p*-value to the model-based *p*-value); however, a close match between the model-based distribution and the permutation distribution (e.g., Figure 10a) provides support for the validity of the model's assumptions.

5.2 | Power analysis

Copula-based models can be used to calculate power, thus to compare multivariate statistical tests under different scenarios. Power calculations require generation of data under a specified alternative hypothesis (H_A). Sliding marginal parameter values (such as μ,

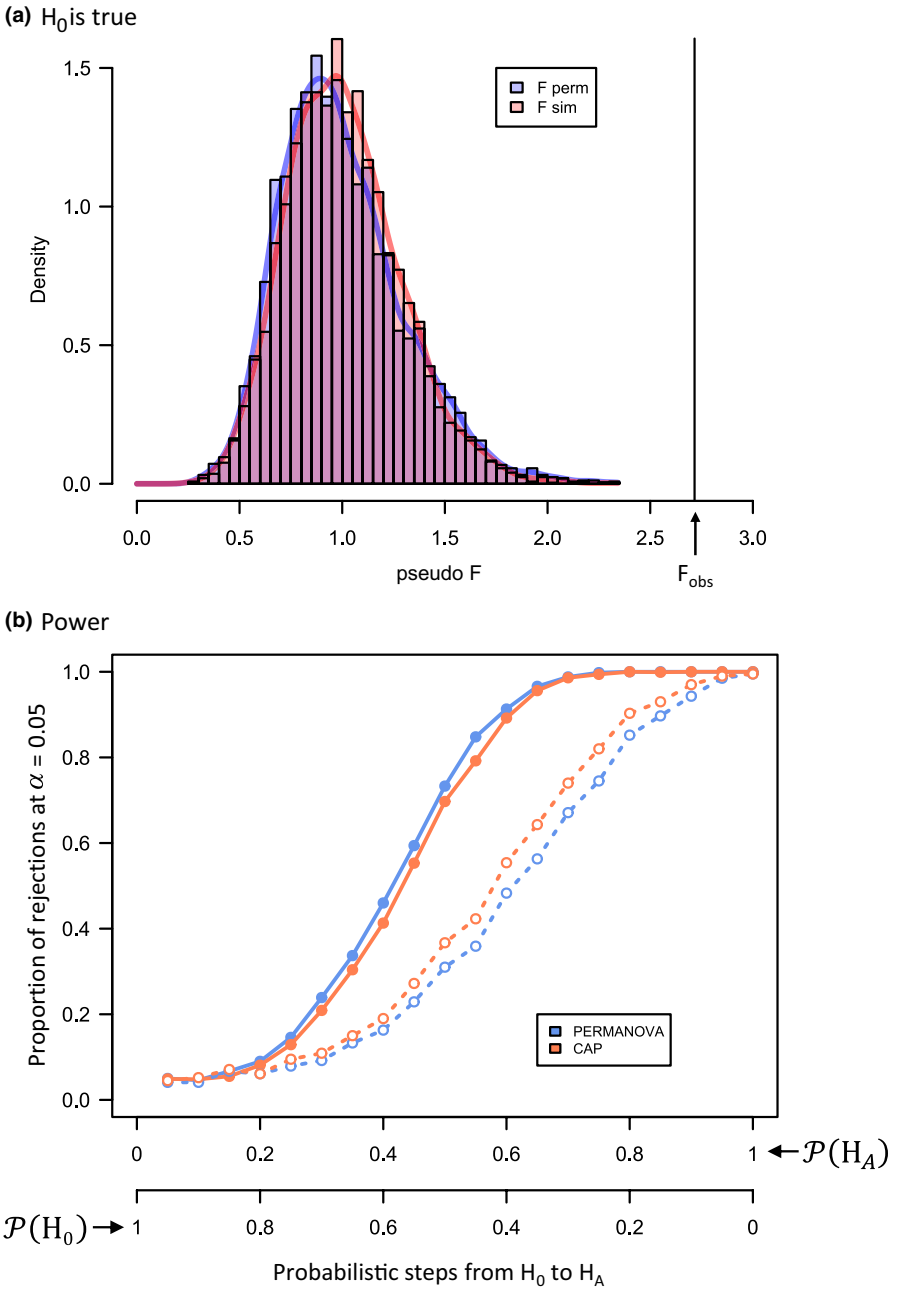


FIGURE 10 (a) Distribution of 5,000 values of the PERMANOVA pseudo-*F* statistic to compare three times of sampling for the Poor Knights dataset based on Bray–Curtis resemblances calculated from square-root-transformed counts of p = 47 fish species obtained under permutation (F perm) or under the copula model (F sim); and (b) empirical power of PERMANOVA or CAP for 1,000 simulated datasets at each of 20 equal steps (as a multinomial mixture of probabilities between H_0 and H_A) for all 47 fish species (filled symbols) or for a subset of 16 fish species only that had estimated copula correlations of $\rho \geq 0.7$ (open symbols). For each simulated dataset, *p*-values were calculated using 999 permutations and seven principal coordinate axes were used for the CAP approach

θ, and/or π) between H_0 and H_A separately for each species can produce unrealistic combinations of parameters. A smooth monotonic power curve is obtained, however, by modeling the continuum between H_0 and H_A as a sliding scale of mixture probabilities.

Let $\mathscr{P}$ denote a $g \times g$ matrix of elements $\mathscr{P}_{ij}$, the probability of drawing a sample $\boldsymbol{y}_{sim}$ for simulated group i (rows) from original group j (columns). For the Poor Knights' dataset, we may consider matrices of probabilities under H_0 ($\mathscr{P}_0$) and under H_A ($\mathscr{P}_A$) as:

$$\mathcal{P}_0 = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \text{ and } \mathcal{P}_A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

respectively (Figure 9b). Next, let f_k be the fractional probabilistic distance from H_0 to H_A in a chosen number of steps (k = 1, …, n_{steps}), beginning with $f_1 = 0$ when H_0 is true and $f_{n_{steps}} = 1$ when H_A is true (Figure 9b). Note that rejecting H_0 is logically distinguishable from the assertion that H_A is true. Note also that a wide variety of H_A may be specified. At a given step k, the probabilities $\mathscr{P}_k$ used to simulate data are as follows:

$$\mathcal{P}_k = (1 - f_k)\,\mathcal{P}_0 + f_k \mathcal{P}_A$$

To provide an example, we generated power curves with n_{steps} = 20 and 1,000 simulated datasets per step for the Poor Knights' dataset and compared the empirical power of PERMANOVA with canonical analysis of principal coordinates (CAP; Anderson & Robinson, 2003; Anderson & Willis, 2003). Tests were done using 999 permutations for each simulated dataset, and CAP analyses were done using seven principal coordinates (which maximized allocation success). When all variables were included, PERMANOVA was more powerful than CAP (Figure 10b); however, when only a subset of fish species having strong associations were included (namely, those having at least one association with an estimated $\rho \geq 0.70$ in the copula model; there were 16 of these), then CAP was more powerful than PERMANOVA (Figure 10b). As an aside, we noted that generating power curves using a slightly different null hypothesis (i.e., where data under H_0 were generated from a model having a single set of parameters estimated from the full set of N sampling units rather than being a multinomial mixture of equiprobable draws from three groups having three separate sets of parameters) made no substantive difference to any of the above results.

6 | DISCUSSION

Copulas provide a rich and flexible approach for modeling associations among disparate types of variables. Recent advances in statistical methods to estimate parameters for copulas having discrete marginal distributions using MCEM (see Appendix 1 below) open new doors for modeling count data. By allowing marginal and copula parameters to vary over time, we uncovered a striking increase in the strengths of associations among fish species after the cessation of fishing at a no-take marine reserve (Figure 5). Naturally, the ecological mechanisms responsible for generating these associations cannot be inferred from observational data alone, but would require additional investigations.

Generalized linear models, GLLVMs, and GJAMs all have tremendous potential for capably modeling count data, particularly if they are extended to allow for changes in over-dispersion or zero-inflation within a species, and changes in correlations among species in different habitats. They do, however, have a few natural limitations. GJAMs avoid using classical statistical distributions, but this comes at a cost—the utility of explicit count distributions for characterizing individual species as univariate variables is lost, and how to choose partition widths to accommodate different mean–variance relationships remains unclear. In GLLVMs, the relationship between each species and each latent variable is effectively linear on a log-scale (when the default log link for count data is used), which may or may not be appropriate/desirable. Also, the latent variable mechanism for inducing correlations will affect estimation of individual species' over-dispersions (and vice versa), making these two conceptually distinct features difficult to disentangle.

Copula models can be readily extended to include other types of variables commonly encountered in ecology, such as biomass, percentage cover, ordinal data, or mixtures of these with counts. They share some of the desirable features of GLLVMs and GJAMs while presenting some distinct advantages. All three approaches can use an underlying MVN distribution to model associations, but copulas can also use other association models, accommodate a wider variety of parametric marginal distributions than GLMs, and do not entangle the association model with the marginal model. Although outside the scope of the present study, a natural next step would be to explore the predictive capabilities of GLLVMs, GJAMs and copula models across a broad range of ecological datasets.

An ecologically meaningful index of association between species (excluding joint absences) is well-preserved by a Gaussian copula model. This has obvious immediate advantages, as methods to achieve parsimony in MVN models abound (Huang & Chen, 2015; Popovic et al., 2018). However, other types of copulas (Genest & Favre, 2007; Schölzel & Friederichs, 2008), including pair-copula constructions (such as vine copulas, see Aas, Czado, Frigessi, & Bakken, 2009; Bedford & Cooke, 2001, 2002; Brechmann & Schepsmeier, 2013), and also non-parametric methods (Iman & Conover, 1982), all deserve further exploration for their potential use in ecology.

The identification of appropriate marginal distributions for modeling abundances also deserves more study. For fish assemblages at the Poor Knights, changes in over-dispersion and zero-inflation through time were clearly evident (Supporting Information Table S2), highlighting the need for flexibility beyond typical exponential families used in GLMs. The relationship $\sigma^2 = \alpha\mu^\beta$ ("Taylor's power law") is virtually ubiquitous for counts of any organism, with α and β being species-specific (Kendal, 2004; Taylor, Woiwood, & Perry, 1978). Variance–mean relationships that follow a power law can be modeled using Tweedie distributions (Tweedie, 1984), a subset of

exponential dispersion models (Jørgensen, 1997) or using contagious distributions (Douglas, 1980; Neyman, 1939) under a generalized Poisson model (Clarke, Chapman et al., 2006; Coly, Yao, Abrial, & Charras-Carrido, 2016). Special cases of the generalized Poisson include the NB (Quenouille, 1949), Neyman Type A (Neyman, 1939), Pólya-Aeppli (Kendall & Stuart, 1963), the discretized Poisson-gamma (Foster & Bravington, 2013; Kendal, 2004), and the Poisson lognormal (Aitchison & Ho, 1989; Preston, 1948). Under-dispersion, where $\sigma^2 < \mu$ (Rogers, 1974), which can occur for organisms exhibiting territorial behavior, or allelopathy (Rice, 1984), is also understudied. Copulas allow *any* marginal distributions to be used for individual species, including with zero-inflation, facilitating broader and deeper investigations of this topic.

A fundamental question remains: What are the limits of our approach? For multivariate data, the available degrees of freedom (*df*), provided sampling units are independent, are likely to be bounded such that $N \leq df < (N \times p)$, and to depend on the level of species' inter-associations. One alternative to our proposed preliminary screening for significant pair-wise associations would be to identify subsets consisting of coherent groups of associated species (Somerfield & Clarke, 2013). These practical permutation-based approaches may be complimented (or replaced) by direct regularization/shrinkage of the copula covariance matrix (Schäfer & Strimmer, 2005). One might also consider joint estimation of copula and marginal parameters. In any case, how to assess parsimony/model complexity in the full framework is an open question.

We have focussed on an ANOVA-type study design. We chose to fit separate parameters (copula plus marginals) to data from each group. However, sampling units might, instead, occur along one or more measured environmental gradients. Continuous predictors can be included in marginal distributions for each variable (as in GLMs; Warton et al., 2015, Niku et al., 2017; Popovic et al., 2018). However, responses of species to gradients are generally unimodal and may be modeled this way (Jamil & ter Braak, 2013; Yee, 2004, 2015), either along each margin or potentially inside the Gaussian copula space. Associations would remain constant using such an approach; however, copula correlations themselves might also be modeled as a function of environmental variables (Nikoloulopoulos & Karlis, 2010)—an idea worth pursuing.

Our approach catered well to varying zero-inflation, but did not optimize models of rare species. Rare taxa are difficult to model (Elith et al., 2006; Fithian, Elith, Hastie, & Keith, 2015) and may not occur randomly; some sites harbor greater coincidences of singletons (Ellingsen, Hewitt, & Thrush, 2007). SDMs can fail to capture the nature of inter-specific associations reliably, particularly for organisms having low probabilities of occurrence (Zurell et al., 2018). Observational data are often too sparse to model rare taxa well individually, but richness (number of species per sampling unit) can be modeled as a Poisson (or Poisson-binomial) random variable (Calabrese, Certain, Kraan, & Dormann, 2014; Gavish et al., 2017). Thus, future model developments could include richness as an additional response variable in a multivariate copula. Relationships between richness and abundances of prevalent species (or environmental variables) could be estimated, allowing potential clustering of rare taxa.

The proposed analysis pathway enables researchers to achieve a greater understanding of the roles and relationships among individual species, as well as providing a novel approach to ordination and power analysis for investigating community-level hypotheses. A unique feature of this framework is that we do not consider model-based methods (such as GLMs, GLLVMs, GJAMs, or copulas) as running counter to dissimilarity-based methods (such as ANOSIM, MDS, PERMANOVA or CAP). Rather, they are complementary: It is not a case of "either, or," but a case of "yes, and…." Probabilistic statistical models are essential for characterizing assemblages on a per-species basis, including estimation of useful interpretable parameters (e.g., Supporting Information Table S2, Figure 5), and also for simulation and prediction. Added value clearly attends the casting of simulations from joint-species models into dissimilarity spaces. Dissimilarity-based methods integrate information across all species in a way that individual species-based models do not. Fundamental ecological concepts such as proportions of species shared, turnover, beta diversity, variation in identities of species, or gestalt shifts in composition are all readily examined through the use of meaningful resemblance measures (Anderson et al., 2011; Anderson, Ellingsen, & McArdle, 2006; Clarke, Somerfield et al., 2006; Kraft et al., 2011; Legendre & De Cáceres, 2013). Ordinations that show not only relationships among centroids but also probabilistic variability in centroid positions (e.g., Figures 7a and 8a) are highly desirable. Moreover, the behavior of dissimilarity-based tests, historically prized for their broad utility and lack of assumptions, can now be further explored under carefully formulated hypotheses articulated by formal joint statistical models (Figures 9b and 10b). By using the latest model-based approaches in tandem with evolving community-level approaches, as proposed here, we can draw the best from both worlds.

We consider that copula-based joint models of species count data, particularly when combined with dissimilarity-based tools, provide a rich new suite of flexible methods that will generate many new scientific insights in the analysis of ecological communities.

ACKNOWLEDGMENTS

MJA was generously supported by a James Cook Fellowship from the Royal Society of New Zealand. AEP was supported by a travel grant from PRIMER-e (Quest Research Limited).

AUTHOR CONTRIBUTIONS

MJA and PD conceived the ideas, PD contributed mathematical formulations for the MCEM and wrote Appendix 1, AP provided R code with contributions from PD and MJA, AEM proposed the idea of mixture models for power analyses, and MJA wrote the manuscript. All authors approved the final draft of the manuscript.

DATA ACCESSIBILITY

Data consisting of counts of abundances of fishes from the Poor Knights Islands are provided as Supporting Information Table S1 and are also available from Dryad Digital Repository: https://doi.org/10.5061/dryad.3s6rm0f.

ORCID

Marti J. Anderson https://orcid.org/0000-0002-4018-4049

REFERENCES

Aas, K., Czado, C., Frigessi, A., & Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, *44*(2), 182–198. https://doi.org/10.1016/j.insmatheco.2007.02.001

Aitchison, J., & Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, *76*, 643–653. https://doi.org/10.1093/biomet/76.4.643

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

Anderson, M. J. (2017). *Permutational multivariate analysis of variance (PERMANOVA)*. Wiley StatsRef: Statistics Reference Online. 1–15. Article ID: stat07841. https://doi.org/10.1002/9781118445112.stat07841

Anderson, M. J., Ellingsen, K. E., & McArdle, B. H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecology Letters*, *9*, 683–693. https://doi.org/10.1111/j.1461-0248.2006.00926.x

Anderson, M. J., & Robinson, J. (2003). Generalized discriminant analysis based on distances. *Australian New Zealand Journal of Statistics*, *45*, 301–318. https://doi.org/10.1111/1467-842X.00285

Anderson, M. J., & Willis, T. J. (2003). Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology. *Ecology*, *84*, 511–525. https://doi.org/10.1890/0012-9658(2003)084[0511:CAOPCA]2.0.CO;2

Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., … Swenson, N. G. (2011). Navigating the multiple meanings of β diversity: A roadmap for the practicing ecologist. *Ecology Letters*, *14*, 19–28. https://doi.org/10.1111/j.1461-0248.2010.01552.x

Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, *19*, 134–143. https://doi.org/10.1111/j.1466-8238.2009.00490.x

Bedford, T., & Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, *32*, 245–268.

Bedford, T., & Cooke, R. M. (2002). Vines – a new graphical model for dependent random variables. *Annals of Statistics*, *30*, 1031–1068. https://doi.org/10.1214/aos/1031689016

Bliss, C. I., & Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data and note on the efficient fitting of the negative binomial. *Biometrics*, *9*, 176–200. https://doi.org/10.2307/3001850

Booth, J. G., & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B*, *61*, 265–285. https://doi.org/10.1111/1467-9868.00176

Brechmann, E. C., & Schepsmeier, U. (2013). Modeling dependence with C- and D-Vine copulas: The R package CDVine. *Journal of Statistical Software*, *52*(3), 1–27.

Caffo, B. S., Jank, W., & Jones, G. L. (2005). Ascent-based Monte Carlo expectation- maximization. *Journal of the Royal Statistical Society: Series B*, *67*, 235–251. https://doi.org/10.1111/j.1467-9868.2005.00499.x

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, *23*, 99–112. https://doi.org/10.1111/geb.12102

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., & West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of American Statistical Association*, *103*, 1438–1456. https://doi.org/10.1198/016214508000000869

Chase, J. M., Abrams, P. A., Grover, J. P., Diehl, S., Chesson, P., Holt, R. D., … Case, T. J. (2002). The interaction between predation and competition: A review and synthesis. *Ecology Letters*, *5*, 302–315. https://doi.org/10.1046/j.1461-0248.2002.00315.x

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. J., & Zhang, S. (2017). Generalized joint attribute modelling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs*, *87*, 34–56.

Clark, N. J., Wells, K., & Lindberg, O. (2018). Unravelling changing interspecific interactions across environmental gradients using Markov random fields. *Ecology*, *99*, 1277–1283. https://doi.org/10.1002/ecy.2221

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, *18*, 117–143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x

Clarke, K. R., Chapman, M. G., Somerfield, P. J., & Needham, H. R. (2006). Dispersion-based weighting of species counts in assemblage analyses. *Marine Ecology Progress Series*, *320*, 11–27.

Clarke, K. R., Somerfield, P. J., & Chapman, M. G. (2006). On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis measure for denuded assemblages. *Journal of Experimental Marine Biology and Ecology*, *330*, 55–80.

Coly, S., Yao, A.-F., Abrial, D., & Charras-Carrido, M. (2016). Distributions to model overdispersed count data. *Journal De La Société Française De Statistique*, *157*, 39–63.

Dauwels, J., Yu, H., Xu, S., & Wang, X. (2013). *Copula Gaussian graphical model for discrete data*. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6283–6287). Vancouver, BC.

De Valpine, P., Scranton, K., Knape, J., Ram, K., & Mills, N. J. (2014). The importance of individual developmental variation in stage-structured population models. *Ecology Letters*, *17*, 1026–1038. https://doi.org/10.1111/ele.12290

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

Douglas, J. B. (1980). *Analysis with standard contagious distributions*. Burtonsville, MD: International Co-operative Publishing House.

Dunstan, P. K., Foster, S. D., & Darnell, R. (2011). Model based grouping of species across environmental gradients. *Ecological Modelling*, *22*, 955–963. https://doi.org/10.1016/j.ecolmodel.2010.11.030

Dunstan, P. K., Foster, S. D., Hui, F. K. C., & 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.

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software*, *21*, 7. http://www.jstatsoft.org/v21/i07

Dwyer, P. S. (1967). Some applications of matrix derivatives in multivariate analysis. *Journal of American Statistical Association*, *62*, 607–625. https://doi.org/10.1080/01621459.1967.10482934

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Journal of Agricultural, Biological, and Environmental Statistics*, *40*, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Ellingsen, K. E., Hewitt, J. E., & Thrush, S. F. (2007). Rare species, habitat diversity and functional redundancy in marine benthos. *Journal of Sea Research*, *58*, 291–301. https://doi.org/10.1016/j.seares.2007.10.001

Faugeras, O. P. (2017). Inference for copula modeling of discrete data: A cautionary tale and some facts. *Dependence Modeling*, *5*, 121–132. https://doi.org/10.1515/demo-2017-0008

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*, 424–438. https://doi.org/10.1111/2041-210X.12242

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

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*, 432–441. https://doi.org/10.1093/biostatistics/kxm045

Gavish, Y., Marsh, C. J., Kuemmerlen, M., Stoll, S., Haase, P., & Kunin, W. E. (2017). Accounting for biotic interactions through alpha-diversity constraints in stacked species distribution models. *Methods in Ecology and Evolution*, *8*, 1092–1102. https://doi.org/10.1111/2041-210X.12731

Genest, C., & Favre, A.-C. (2007). Everything you always wanted to know about copula modelling but were afraid to ask. *Journal of Hydrologic Engineering*, *12*, 347–368.

Genest, C., & Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin*, *37*, 475–515. https://doi.org/10.1017/S0515036100014963

Goldberg, D. E., & Landa, K. (1991). Competitive effect and response: Hierarchies and correlated traits in the early stages of competition. *Journal of Ecology*, *79*, 1013–1030. https://doi.org/10.2307/2261095

Golding, N., & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, *7*, 598–608. https://doi.org/10.1111/2041-210X.12523

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*, 857–871.

Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, *6*, 465–473. https://doi.org/10.1111/2041-210X.12332

Harris, D. J. (2016). Inferring species interactions from co-occurrence data with Markov networks. *Ecology*, *97*, 3308–3314. https://doi.org/10.1002/ecy.1605

Huang, F., & Chen, S. (2015). Joint learning of multiple sparse matrix Gaussian graphical models. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(11), 2606–2620. https://doi.org/10.1109/TNNLS.2014.2384201

Hui, F. K. C. (2016). BORAL – Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods in Ecology and Evolution*, *7*, 744–750.

Hui, F. K. C., 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. https://doi.org/10.1111/2041-210X.12236

Iman, R. L., & Conover, W. J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation*, *11*(3), 311–334. https://doi.org/10.1080/03610918208812265

Jamil, T., & ter Braak, C. J. F. (2013). Generalized linear mixed models can detect unimodal species environment relationships. *PeerJ*, *1*, e95. https://doi.org/10.7717/peerj.95: 1-14

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, *94*, 401–419. https://doi.org/10.1016/j.jmva.2004.06.003

Jørgensen, B. (1997). *The theory of exponential dispersion models*. London, UK: Chapman & Hall.

Kendal, W. S. (2004). Taylor's ecological power law as a consequence of scale invariant exponential dispersion models. *Ecological Complexity*, *1*, 193–209. https://doi.org/10.1016/j.ecocom.2004.05.001

Kendall, B. E., Bjørnstad, O. N., Bascompte, J., Keitt, T. H., & Fagan, W. F. (2000). Dispersal, environmental correlation, and spatial synchrony in population dynamics. *American Naturalist*, *155*, 628–636. https://doi.org/10.1086/303350

Kendall, M. G., & Stuart, A. (1963). *The advanced theory of statistics*, Vol. *I*, 2nd ed. London, UK: Griffin.

Kraft, N. J., Comita, L. S., Chase, J. M., Sanders, N. J., Swenson, N. G., Crist, T. O., … Myers, J. A. (2011). Disentangling the drivers of beta-diversity along latitudinal and elevational gradients. *Science*, *333*(6050), 1755–1758.

Legendre, P., & De Cáceres, M. (2013). Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. *Ecology Letters*, *16*, 951–963. https://doi.org/10.1111/ele.12141

Legendre, P., & Legendre, L. (2012). *Numerical ecology*, 3rd, English ed. Amsterdam, the Netherlands: Elsevier Science BV.

Mai, J.-F., & Scherer, M. (2017). *Simulating copulas: Stochastic models, sampling algorithms, and applications*, 2nd ed. Singapore, Singapore: World Scientific Publishing.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, T. A. J., & Possingham, H. P. (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters*, *8*, 235–1246. https://doi.org/10.1111/j.1461-0248.2005.00826.x

McArdle, B. H., & Anderson, M. J. (2004). Variance heterogeneity, transformations and models of species abundance: A cautionary tale. *Canadian Journal of Fisheries and Aquatic Science*, *61*, 1294–1302. https://doi.org/10.1139/f04-051

McArdle, B. H., Gaston, K. J., & Lawton, J. H. (1990). Variation in the size of animal populations: Patterns, problems and artefacts. *Journal of Animal Ecology*, *59*, 439–454. https://doi.org/10.2307/4873

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., … White, E. P. (2007). Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, *10*, 995–1015. https://doi.org/10.1111/j.1461-0248.2007.01094.x

Meng, Z., Eriksson, B., & Hero, A. (2014). *Learning latent variable Gaussian graphical models*. In T. Jebara &E. P. Xing (Eds.), *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (pp. 1269–1277) JMLR Workshop and Conference Proceedings.

Neyman, J. (1939). On a new class of "contagious" distributions applicable in entomology and bacteriology. *The Annals of Mathematical Statistics*, *10*, 35–57. https://doi.org/10.1214/aoms/1177732245

Nieto-Lugilde, D., Maguire, K. C., Blois, J. L., Williams, J. W., & Fitzpatrick, M. C. (2017). Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models. *Methods in Ecology and Evolution*, *9*, 834–848. https://doi.org/10.1111/2041-210X.12936

Nikoloulopoulos, A. K., & Karlis, D. (2009). Modeling multivariate count data using copulas. *Communications in Statistics - Simulation and Computation*, *39*, 172–187. https://doi.org/10.1080/03610910903391262

Nikoloulopoulos, A. K., & Karlis, D. (2010). Regression in a copula model for bivariate count data. *Journal of Applied Statistics*, *37*, 1555–1568. https://doi.org/10.1080/02664760903093591

Niku, J., Warton, D. I., Hui, F. K. C., & Taskinen, S. (2017). Generalized linear latent variable models for multivariate count and biomass data in

ecology. *Journal of Agricultural, Biological and Environmental Statistics*, *22*, 498–522. https://doi.org/10.1007/s13253-017-0304-7

Ovaskainen, O., Abrego, N., Halme, P., & Dunson, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, *7*, 549–555.

Ovaskainen, O., Roy, D. B., Fox, R., & Anderson, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*, *7*, 428–436.

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., … Abrego, N. (2017). How to make more out f community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, *20*, 561–576.

Pacala, S. W., & Roughgarden, J. (1985). Population experiments with the *Anolis* lizards of St. Maarten and St. Eustatius. *Ecology*, *66*, 129–141.

Paradis, E., & Claude, J. (2002). Analysis of comparative data using generalized estimating equations. *Journal of Theoretical Biology*, *218*, 175–185. https://doi.org/10.1006/jtbi.2002.3066

Perry, M. A., Mitchell, R. J., Zutter, B. R., Glover, G. R., & Gjerstad, D. H. (1994). Seasonal variation in competitive effect on water stress and pine responses. *Canadian Journal of Forest Research*, *24*, 1440–1449. https://doi.org/10.1139/x94-186

Pinheiro, J. C., & Bates, D. M. (1996). Unconstrained parameterizations for variance-covariance matrices. *Statistics and Computing*, *6*, 289–296.

Pollock, K. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., … 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. https://doi.org/10.1111/2041-210X.12180

Popovic, G. C., Hui, F. K. C., & Warton, D. I. (2018). A general algorithm for covariance modeling of discrete data. *Journal of Multivariate Analysis*, *165*, 86–100. https://doi.org/10.1016/j.jmva.2017.12.002

Preston, F. W. (1948). The commonness, and rarity, of species. *Ecology*, *29*, 254–283. https://doi.org/10.2307/1930989

Quenouille, M. M. (1949). A relation between the logarithmic, Poisson and negative binomial series. *Biometrics*, *5*, 162–164. https://doi.org/10.2307/3001917

Rice, E. L. (1984). *Allelopathy*, 2nd ed. Orlando, FL: Academic Press Inc.

Rogers, A. (1974). *Statistical analysis of spatial dispersion: The quadrat method*. London, UK: Pion Limited.

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*, 1175–1189. https://doi.org/10.2202/1544-6115.1175

Schölzel, C., & Friederichs, P. (2008). Multivariate non-normally distributed random variables in climate research – introduction to the copula approach. *Nonlinear Processes in Geophysics*, *15*, 761–772.

Shi, P., & Valdez, E. A. (2014). Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics*, *55*, 18–29. https://doi.org/10.1016/j.insmatheco.2013.11.011

Sklar, A. (1959). Fonctions de répartition à *n* dimensions et leurs marges. *Publications De L'institut De Statistique De L'université De Paris*, *8*, 229–231.

Smith, A. N. H., Anderson, M. J., & Millar, R. B. (2012). Incorporating the intraspecific occupancy-abundance relationship into zero-inflated models. *Ecology*, *93*, 2526–2532. https://doi.org/10.1890/12-0460.1

Somerfield, P. J., & Clarke, K. R. (2013). Inverse analysis in non-parametric multivariate analyses: Distinguishing groups of associated species which covary coherently across samples. *Journal of Experimental Marine Biology and Ecology*, *449*, 261–273. https://doi.org/10.1016/j.jembe.2013.10.002

Taylor, L. R. (1961). Aggregation, variance and the mean. *Nature*, *189*, 732–735. https://doi.org/10.1038/189732a0

Taylor, L. R., Woiwood, I. P., & Perry, J. N. (1978). The density-dependence of spatial behaviour and the rarity of randomness. *Journal of Animal Ecology*, *47*, 383–406. https://doi.org/10.2307/3790

Taylor, L. R., Woiwood, I. P., & Perry, J. N. (1979). The negative binomial as a dynamic ecological model for aggregation, and the density dependence of k. *Journal of Animal Ecology*, *48*, 289–304. https://doi.org/10.2307/4114

ter Braak, C. J. F. (1996). *Unimodal models to relate species to environment*. Wageningen, the Netherlands: DLO-Agricultural Mathematics Group.

Thorson, J. T., Ianelli, J. N., Larsen, E. A., Ries, L., Scheuerell, M. D., Szuwalski, C., & Zipkin, E. F. (2016). Joint dynamic species distribution models: A tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, *25*, 1144–1158. https://doi.org/10.1111/geb.12464

Thorson, J. T., Scheuerell, M. D., Shelton, A. O., See, K. E., Skaug, H. J., & Kristensen, K. (2015). Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution*, *6*, 627–637. https://doi.org/10.1111/2041-210X.12359

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In J. K. Ghosh & J. Roy (Eds.), *Statistics: Applications and new directions* (pp. 579–604). Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Calcutta, India: Indian Statistical Institute.

Ullah, I., & Jones, M. B. (2015). Regularised MANOVA for high-dimensional data. *Australian and New Zealand Journal of Statistics*, *57*, 377–389. https://doi.org/10.1111/anzs.12126

Vellend, M. (2001). Do commonly used indices of ↕-diversity measure species turnover? *Journal of Vegetation Science*, *12*, 545–552.

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. (2011). Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics*, *67*, 116–123. https://doi.org/10.1111/j.1541-0420.2010.01438.x

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, 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. https://doi.org/10.1016/j.tree.2015.09.007

Warton, D. I., & Guttorp, P. (2011). Compositional analysis of over-dispersed counts using generalized estimating equations. *Environmental and Ecological Statistics*, *18*, 427–446. https://doi.org/10.1007/s10651-010-0145-9

Wei, G. C. G., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of American Statistical Association*, *85*, 699–704. https://doi.org/10.1080/01621459.1990.10474930

Welsh, A. H., Cunningham, R. B., Donnelly, C. F., & Lindenmayer, D. B. (1996). Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling*, *88*, 297–308. https://doi.org/10.1016/0304-3800(95)00113-1

Wenger, S. J., & Freeman, M. C. (2008). Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology*, *89*, 2953–2959. https://doi.org/10.1890/07-1127.1

Wheldon, M. C., Anderson, M. J., & Johnson, B. W. (2007). Identifying treatment effects in multi-channel measurements in electroencephalographic studies: Multivariate permutation tests and multiple comparisons. *Australian and New Zealand Journal of Statistics*, *49*, 397–413. https://doi.org/10.1111/j.1467-842X.2007.00491.x

White, G. C., & Bennetts, R. E. (1996). Analysis of frequency count data using the negative binomial distribution. *Ecology*, *77*, 2549–2557. https://doi.org/10.2307/2265753

Whittaker, R. H. (1952). A study of summer foliage insect communities in the Great Smoky Mountains. *Ecological Monographs*, *22*, 1–44. https://doi.org/10.2307/1948527

Willis, T. J., & Denny, C. M. (2000). *Effects of Poor Knights Islands Marine Reserve on demersal fish populations*. Report to the Department of Conservation, Research Grant No. 2519. New Zealand, Auckland: Leigh Marine Laboratory, University of Auckland.

Yee, T. W. (2004). A new technique for maximum-likelihood canonical Gaussian ordination. *Ecological Monographs*, *74*, 685–701. https://doi.org/10.1890/03-0078

Yee, T. W. (2015). *Vector generalized linear and additive models: With an implementation in R*. New York, NY: Springer.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, *94*, 19–35. https://doi.org/10.1093/biomet/asm018

Zurell, D., Pollock, L. J., & Thuiller, W. (2018). Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogeneous environments? *Ecography*, *41*, 1–8.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Anderson MJ, de Valpine P, Punnett A, Miller AE. A pathway for multivariate analysis of ecological communities using copulas. *Ecol Evol*. 2019;9:3276–3294. https://doi.org/10.1002/ece3.4948

APPENDIX 1

This appendix outlines the core idea of using a Monte Carlo Expectation Maximization (MCEM, Wei & Tanner, 1990) algorithm to estimate the covariance matrix for a Gaussian copula with discrete marginal distributions that is constrained to have unit variances along the diagonal (hence takes the form of a correlation matrix). By constraining the Gaussian copula covariance matrix in this way, our algorithm ensures that the variances of individual variables remain precisely as specified by their marginal probability mass functions. For other examples of Gaussian copulas for discrete data, see Dauwels, Yu, Xu, and Wang (2013) and, more recently, Popovic et al. (2018), who describe a general covariance modeling framework, including latent variable graphical models (Meng, Eriksson, & Hero, 2014) and sparse factor analysis (Carvalho et al., 2008).

BASIC MCEM

The general setup for MCEM is as follows (see Dempster, Laird, & Rubin, 1977; Wei & Tanner, 1990). Let y be an observed data vector. Let x be latent states or "missing data." Let θ be parameters. Define $f(y|x;\theta)$ and $f(x;\theta)$ to be the probability density (or mass) functions of y given x and x, respectively, as indicated by the arguments. The likelihood to maximize is as follows:

$$L(\theta) = \int f(y|x;\theta) f(x;\theta)\, d($$

Define the maximum likelihood estimate we seek as

$$\hat{\theta} = \arg\max L(\theta).$$

The MCEM algorithm works as follows:

1. Start with some initial value $\theta_{k=0}$ of θ.
2. Draw a sample of many values $X = \{x_i\}$, $i = 1,\ldots, m$, from $f(x|y;\theta_k)$.
3. Find θ_{k+1} as:

$$\theta_{k+1} = \arg\max \frac{1}{m}\sum_{i=1}^{m} \log\left(f(y|x_i;\theta) f(x_i;\theta)\right)$$

The Monte Carlo average is the approximation of:

$$\frac{1}{m}\sum_{i=1}^{m} \log\left(f(y|x_i;\theta) f(x_i;\theta)\right) \approx E_{xy,\theta}\left[\log\left(f(y|x;\theta) f(x;\theta)\right)\right]$$

1. Repeat the previous two steps, which are known as the "*E*"xpectation step and the "*M*"aximization step.

Without the Monte Carlo approximation, it can be proven that, provided θ_k is not a stationary point, iterations will always yield $L(\theta_{k+1}) > L(\theta_k)$, so that θ_k converges to $\hat{\theta}$, unless there are local maxima or other stationary points. With the Monte Carlo approximation, iterations will not converge to a single value, but will instead oscillate about $\hat{\theta}$, with a precision that depends on the number of Monte Carlo samples, m. Thus, the value of m should be determined by the required precision of the results. Operationally, the beauty of MCEM is that we maximize a simple average of log probabilities at each step. This is much easier to work with than the (log of the) expected value of the probabilities from the latent states.

MCEM FOR DISCRETE GAUSSIAN COPULAS

In what follows, we shall think of each discrete observation as having an unobserved fractional component. Let $y_i = (y_{i1}, \ldots, y_{ip})$ be counts of p species from sampling unit (or "site") i. Subscript i will be dropped in discussing the likelihood contribution for one site's data only. Let $f_j(y_j)$ and $F_j(y_j)$ be the marginal probability mass function and cumulative distribution function, respectively, for species j, assuming y_j takes non-negative integer values. Dependence on parameters, θ, is implicit. Let $f(y) = \prod f_j(y_j)$ and $F(y) = [F_1(y_1), \ldots, F_p(y_p)]$. Let $g(y)$ be the joint density function defined using a copula. For a Gaussian copula, if the observations were continuous and f_j's were PDFs, this would be:

$$g(y) = \frac{\phi_{\Sigma}(z)}{\phi_{I}(z)} f(y)$$

where $\phi_{\mathbf{\Sigma}}$ is the multivariate normal density with zero means and unit variances and correlation matrix $\mathbf{\Sigma}$, $\mathbf{I}$ is the identity matrix, and $z = (z_1, \ldots, z_p)$ is defined by $z_j = \Phi^{-1}\left(F_j\left(y_j\right)\right)$, $j = 1,\ldots, p$ where Φ is the standard normal CDF.

Let $\tilde{f}_j\left(y_j\right)$ be a probability density function defined by spreading the probability mass $f_j(y_j)$ uniformly in the interval $[y_j, y_j + 1)$; that means:

$$\tilde{f}(y) = f(\lfloor y \rfloor)$$

Now the discrete copula likelihood can be written as the following integral over a hypercube:

$$\mathcal{L}(\mathbf{\Sigma}|y) = \prod_i \int_{[0,1]^p} \frac{\phi_{\mathbf{\Sigma}}\left(\Phi^{-1}\left(\tilde{F}(y+u)\right)\right)}{\phi_{\mathbf{I}}\left(\Phi^{-1}\left(\tilde{F}(y+u)\right)\right)} \tilde{f}(y+u)\, du$$

where i is for sites, and the integral is over $u \in [0,1]^p$, where p is the number of dimensions.

To work in the space of the multivariate normal variables defined by z, we change variables to $z = \Phi^{-1}\left(\tilde{F}(y+u)\right)$, so $y + u = \tilde{F}^{-1}(\Phi(z))$ and

$$du = \frac{1}{\tilde{f}\left(\tilde{F}^{-1}(\Phi(z))\right)} \phi(z)\, dz.$$

This yields

$$\mathcal{L}(\mathbf{\Sigma}|y) = \prod_i \int \phi_{\mathbf{\Sigma}}(z)\, dz$$

where the integral is over $z \in \Phi^{-1}\left(\tilde{F}(y+u)\right)$ for $u \in [0,1]^p$. That is the region of z that corresponds to $y + u$ falling in the interval $[y_j, y_j + 1)$.

Another way to write this is by integrating over the entire range of z and using an indicator function to exclude z outside the range given above. This is written as

$$\mathcal{L}(\mathbf{\Sigma}|y) = \prod_i \int_z \phi_{\mathbf{\Sigma}}(z)\, I_{\tilde{F}^{-1}(\Phi(z)) \in [y,\, y+1)} dz$$

where $I_{\text{condition}}$ is 1 if "condition" is true, and 0 otherwise. Now we are ready to view this in the MCEM framework:

1. The "latent variable" is z, whose probability density is $\phi_{\mathbf{\Sigma}}(z)$;
2. The parameters are the elements of $\mathbf{\Sigma}$;
3. The "observed data" are the values of z falling in the valid range defined above. The probability of the observed data given z is $I_{\tilde{F}^{-1}(\Phi(z)) \in [y,\, y+1)}$;
4. We consider the marginal distribution parameters for f as given.

With these interpretations, $f(z|y; \mathbf{\Sigma})$ (in the role of $f(x|y; \theta)$ in the section sub-titled "Basic MCEM" above) can be sampled easily. For example, if we need to sample from z given $z \in [a, b)$, we can make use of $\phi_{\mathbf{\Sigma}}$ in the relevant range. Then the MCEM algorithm would be:

1. Start with some initial value $\mathbf{\Sigma}_{k=0}$ of $\mathbf{\Sigma}$.
2. Draw a sample of many values $Z = \{z_i\}$, $i = 1, \ldots, m$, from $f(z|y; \mathbf{\Sigma}_k)$.
3. Find $\mathbf{\Sigma}_{k+1}$ as:

$$\mathbf{\Sigma}_{k+1} = \arg\max \frac{1}{m} \sum_{i=1}^{m} \log\left(\phi_{\mathbf{\Sigma}}\left(z_i\right) I_{\tilde{F}^{-1}(\Phi(z)) \in [y,\, y+1)}\right)$$

Note that the Indicator function will always be 1 (because of how the z_i's were sampled). That leaves only the task of maximizing the log-likelihood of a normal correlation matrix based on the "sample" of z values, given the previous value of the correlation matrix. For a general multivariate normal covariance matrix (having non-unit variances along the diagonal and covariances in off-diagonal elements), we can immediately write down the maximum likelihood estimator (MLE; Dwyer, 1967). However, to obtain the MLE for the correlation matrix, $\mathbf{\Sigma}_{k+1}$, a closed-form expression is not immediately available. We instead use a numerical optimization technique with a spherical parameterization of the correlation matrix (Pinheiro & Bates, 1996, see section 2.3 therein), which allows us to constrain variances to 1, correlations to the range (−1, 1), and ensures a positive-definite result.

Repeat the previous two steps until a stopping criterion has been satisfied. A simple stopping criterion can be implemented by taking a set of the n most recent estimates of the parameters, $\mathbf{\Sigma}_{k-n+1}$, ..., $\mathbf{\Sigma}_k$, calculating the log-likelihood associated with each of those estimates, ℓ_1, ..., ℓ_n, and fitting a simple linear model of ℓ_1, ..., ℓ_n versus the integer values 1, ..., n (corresponding to the n most recent iteration steps in the MC algorithm); the MCEM algorithm is terminated when there is no evidence against the null hypothesis that the slope parameter associated with this linear model is significantly different from zero (H_0: $\beta = 0$). We consider a useful approach (avoiding type II error) will be to require $p > 0.25$ in order to assert that H_0 is true. More efficient stopping criteria may be found by estimating the variance of parameters (Booth & Hobert, 1999) or their log-likelihoods (Caffo, Jank, & Jones, 2005), using just the Monte Carlo samples from a single MCEM iteration.