Submission to: Global Ecology and Biogeography

Article Type: Macroecological Methods

Extension of the Gambin Distribution to Multimodal Species Abundance Distributions

Thomas J. Matthews¹,², Michael K. Borregaard¹, Karl I. Ugland⁴, Colin Gillespie, François Rigal⁵, Rodrigo Ferreira Krüger⁷,⁸, Jon P. Sadler¹, Paulo A.V. Borges², Pedro Cardoso⁶, Rodger J. Whittaker¹,³

¹GEES (School of Geography, Earth and Environmental Sciences), The University of Birmingham, Birmingham, B15 2TT


³Center for Macroecology, Evolution and Climate, Department of Biology, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark

⁴Department of Marine Biology, Institute of Biosciences, University of Oslo, P.O. Box 1066, Blindern, 0316 Oslo, Norway

⁵Environment and Microbiology Team, Université de Pau et des Pays de l’Adour, 9 IPREM UMR CNRS 5254, BP 1155, 64013 Pau Cedex, France

⁶Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland

⁷Conservation Biogeography and Macroecology Group, School of Geography and the Environment, University of Oxford, South Parks Road, Oxford,OX1 3QY, UK

⁸Parasite and Vector Ecology Group, Depto de Microbiologia e Parasitologia, Campus Universitário Capão do Leão, s/nº CEP 96010-900, Pelotas, Rio Grande do Sul, Brasil

*Correspondence: Thomas J. Matthews, School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, B15 2TT, UK

Email: t.j.matthews@bham.ac.uk

Running header: Multimodal gambin distributions

Word count: abstract: words; main text = words; Tables; Figures; references; appendices

Keywords
ABSTRACT

Aim

Innovation

Main conclusions

INTRODUCTION

The species abundance distribution (SAD) has been a core focus of macroecology for over eighty years (e.g. Motomura, 1932; Fisher et al., 1943), and is currently the subject of widespread renewed interest (McGill et al., 2007; Ulrich et al., 2010; McGill, 2011 Borda-de-Água et al., 2017*). Recently, it has been argued that a gamma-binomial (herein ‘gambin’) distribution represents a useful SAD model (Ugland et al. 2007). Gambin is a stochastic model that combines the gamma distribution, in which the scale parameter is fixed at 1, with a binomial sampling method (see Ugland et al. 2007 for a full description of the model). At present, the gambin distribution can only be fitted to data binned into octaves (Xiao et al., In review*). The use of the gamma distribution as the basis of the model provides gambin with substantial flexibility and tests of the gambin model have found that it generally provides a good fit to a wide range of empirical SAD data, often out-competing other candidate SAD models, such as the Poisson lognormal (PLN) and logseries models (Ugland et al. 2007; Matthews et al. 2014a). As the scale parameter is fixed by the value of the largest octave, the unimodal gambin model has a single free parameter (α) which determines the shape of the distribution. Low values of α indicate logseries curve shapes, whilst higher α values indicate more lognormal like curve shapes. Thus, α is an intuitive parameter that has been found to be of use in comparing the SAD of different ecological communities, e.g. disturbed and undisturbed communities, and for testing what variables drive changes in the shape of the SAD along ecological gradients (Dornelas et al. 2011; Matthews & Whittaker, 2015; Picaço et al., 2017; Arellano et al., 2017; Matthews et al., 2017). In sum, gambin provides a flexible SAD model with a single interpretable free parameter that describes the shape of the distribution.

In a recent paper, we derived the maximum likelihood formula for the gambin model and incorporated it within an R package (‘gambin’, version 1.0; Matthews et al. 2014a). However, during this work we observed that gambin was unable to provide a satisfactory fit to empirical SADs which appeared to have multiple modes. This observation is unsurprising given that gambin is a unimodal model. Nonetheless, and despite the potential being recognised since the 1980s (e.g. Ugland & Gray, 1982), it has become increasingly apparent that many empirical SADs are in fact multimodal (Dornelas & Connolly, 2008; Vergnon et
al. 2012; Matthews et al., 2014b; Antão et al., 2017). For example, Antão et al. (2017) found that between 15% and 22% of the 117 empirical SAD datasets they evaluated showed evidence of multimodality, depending on the metric used. Determining whether empirical SADs are multimodal is more than just a search for patterns; rather, multimodal SADs can be indicative of particular processes. For example, the theory of emergent neutrality makes predictions regarding multimodal SADs (Vergnon et al., 2012). Thus, the study of multimodal SAD models can be seen as a priority in SAD research (Antão et al., 2017). Nonetheless, only a few SAD models have been extended to incorporate multiple modes; the main example being the PLN (see Dornelas & Connolly, 2008; see also Tovo et al. In review* for an example using multimodal negative-binomial distributions). Many compound probability distribution models are mathematically and computationally complex and there exists at present limited software packages for fitting multimodal SAD models. For example, fitting multimodal PLN models and getting the optimisation algorithms to converge can be difficult and dependent on the starting parameter values selected. As such, there is a need for a software package that allows researchers to undertake a quick and easy straightforward statistical analysis of multimodality in SAD datasets. A multimodal extension of gambin would be particularly useful in this regard as the model is relatively simple and it would allow for the fit of unimodal and multimodal gambin models to data to be compared analytically using standard statistical methods; for instance, by comparing Akaike’s information criterion values (AIC; Burnham & Anderson, 2002). A multimodal gambin model would also be of use as the \( g \alpha \) parameters (where \( g = \) the number of modes in the distribution) may provide simple metrics for describing the rough shape of the different compound distributions in certain cases.

In this study, we derive the maximum likelihood equations for fitting the bimodal gambin distribution and generalise this approach to fit gambin models with any number \((g)\) of modes. We incorporate these new functions within an updated R package (‘gambin’; version 2.0). We use a mixture of simulations and empirical datasets to test the new models, and examples of how to use the updated package are provided. We do not compare the fit of alternative multimodal SAD models or try and determine the prevalence of multimodal SADs more generally (see Antão et al., 2017) as these aims are beyond the scope of our study, and integrating the gambin binning framework with other compound distributions is not straightforward.

METHODS

The gambin R package (version 2.0)

The original gambin R package (version 1.0; Matthews et al., 2014a) provided functionality to fit the gambin distribution to abundance data using maximum likelihood, estimate \( \alpha \), and calculate confidence intervals, predicted values from the fitted model, goodness of fit, and model comparison metrics (e.g. \( X^2 \), AIC). In addition to providing functions to fit multimodal gambin distributions (described below), the gambin R package (version 2.0) has been updated
to bring it more in line with other distribution functions within the R base ‘stats’ package. For example, the updated gambin package now provides dgambin (probability density function), rgambin (generate random values from a gambin distribution; the returned values relate to a given octave), qgambin (quantile function) and pgambin (cumulative distribution function) functions. The rgambin function is particularly useful as it allows the user to randomly sample from gambin distributions with a given alpha value(s) and number of species. In addition, the fitting optimisation procedure has been improved and the plot function has been integrated with the standard barplot, lines and points functions (R Core Team, 2017) to enable standard plotting arguments to be passed to the plot function, thus improving usability.

As the likelihood optimisation procedure for multimodal gambin models can be time consuming, the updated package provides the option of using parallel processing to speed up optimisation.

### Bimodal gambin distribution

The bimodal gambin model is simply the integration of two gambin distributions. In order to allow for the subdivision of all of the observed objects (species in the context of SADs) $(y_{obs})$, a parameter (split) is needed that describes the fraction of objects belonging to the first distribution. Thus, the number of observed objects is split into two groups, consisting of $split \times y_{obs}$ and $(1 - split) \times y_{obs}$, objects, respectively. Thus, $y_{obs} = (split \times y_{obs}) + ((1 - split) \times y_{obs})$.

With no extra information we may therefore assume that the number of objects in the $k$-th interval $(k = 1, 2, ... , imax)$ are $split^{\alpha_1}$ and $(1-split)^{\alpha_2}$ in the first and second group, respectively. Whilst the max octave of the second group is set to equal the max octave of the empirical distribution, the max octave of the first group can vary. The corresponding likelihood function for this model will therefore contain four parameters: the shape parameters for the first and second group ($\alpha_1$ & $\alpha_2$), the max octave for the first group, and one splitting parameter (split) representing the fraction of objects in the first group. The likelihood function for the bimodal gambin distribution can then be derived as follows:

First, it is necessary to treat each of the two groups separately. For the first group, the ‘success-distribution’ (see Matthews et al., 2014a) is given by $(G_1(\alpha_1), G_2(\alpha_1), ... , G_{99}(\alpha_1), G_{100}(\alpha_1))$ and is calculated using the equations in Matthews et al. (2014a) along with the shape parameter $\alpha_1$ in order to obtain the marginal interval probabilities $(P_{1\alpha_1}(\alpha_1), P_{2\alpha_1}(\alpha_1), ... , P_{max\alpha_1}(\alpha_1))$, given the shape parameter $\alpha_1$, for the first group. Thus, the likelihood function for the first group is given by:

$$L_1(\alpha_1, split) = P_1^{\alpha_1 y_{obs}}(\alpha_1) \prod_{i=2}^{max} P_{i\alpha_1}^{\alpha_1 y_{obs}}(\alpha_1) \prod_{i=max+1}^{99} P_{i\alpha_1}^{\alpha_1 y_{obs}}(\alpha_1)$$

(1)

This procedure is then repeated for the second group where the ‘success-distribution’, $(G_1(\alpha_2), G_2(\alpha_2), ... , G_{99}(\alpha_2), G_{100}(\alpha_2))$, is calculated with the shape parameter $\alpha_2$ in order to obtain the marginal interval probabilities $(P_{1\alpha_2}(\alpha_2), P_{2\alpha_2}(\alpha_2), ... , P_{max\alpha_2}(\alpha_2))$, given the shape parameter $\alpha_2$, for the second group. Thus, the likelihood function for the second group is given by:

Commented [T1]:
Colin – this section, and the likelihood equations in particular, were written when the max octave of both compound distributions were set to equal the empirical max octave. As this has now changed (i.e. the max octave of the first compound distribution is allowed to vary), I have edited the text to state that this now happens, I am not sure how to update the likelihood equations?

This also relates to the next subsection of gambin models with more than two modes.

Commented [T2]:
Colin – in the fit_gambin function file in the R package, the df for the log lik is calculated as:

```
no_of_components^2 + (no_of_components - 1)
```

Which for a bimodal model equals $4 + 1 = 5$, but I can only work out 4: alpha one and two, split and max octave of the first distribution (as max oct of the second is equal to max oct of the empirical distribution)?
The maximum likelihood estimators of the three parameters are those values which maximize the likelihood function:

\[ L(\alpha_1, \alpha_2, \text{split}) = L_1^{(1-\text{split})\alpha_1}(\alpha_1) \cdot L_2^{(1-\text{split})\alpha_2}(\alpha_2) \cdot L_{\text{split}}^{\alpha_2 \cdot \text{split}}(\alpha_2) \quad (2) \]

The corresponding gambin model of the observed frequencies is calculated as the sum of the abundances in the intervals of each group:

\[ \hat{y}_k = \text{split} \cdot \hat{y}_{\text{obs}} \cdot \hat{P}_1(\hat{\alpha}_{1, \text{obs}}) + (1 - \text{split}) \cdot \hat{y}_{\text{obs}} \cdot \hat{P}_1(\hat{\alpha}_{2, \text{obs}}) \]
for \( k = 1, 2, ..., \text{imax} \).

\[ \text{(4)} \]

Gambin distributions with \( g \) modes

It is relatively straightforward to extend the above approach for fitting the bimodal gambin model by maximum likelihood, to fitting gambin models with \( g \) modes. For each additional mode, a further three parameters are needed:

- \( \text{split}_m \) - the fraction of species in the \( m \)-th group for \( m = 1, 2, ..., g \); the final abundance in the last group is calculated as \( \text{split}_g = 1 - \text{split} + \text{split}_1 + ... + \text{split}_{g-1} \)
- \( \alpha_m \) - the \( \alpha \) parameter of each subgroup: \( m = 1, 2, ..., g \)
- \( \text{MaxOct}_m \) - the max octave of the \( m \)-th group

Each group is again treated separately. For the \( m \)-th group the ’success-distribution’, \( \{G_1(\alpha_m), G_2(\alpha_m), ..., G_{99}(\alpha_m), G_{100}(\alpha_m)\} \) is calculated, given the shape parameter \( \alpha_m \), in order to obtain the marginal interval probabilities \( \{P_1(\alpha_m), P_2(\alpha_m), ..., P_{\text{max}}(\alpha_m)\} \). The likelihood function for the \( m \)-th group (\( m = 1, 2, ..., g \)) is given by:

\[ L_m(\alpha_m, \text{split}_m) = P_1^{\alpha_m \cdot \text{split}_m}(\alpha_m) \cdot P_2^{\alpha_m \cdot \text{split}_m}(\alpha_m) \cdot L_{\text{split}}^{\alpha_m \cdot \text{split}}(\alpha_m) \quad (5) \]

The maximum likelihood estimators of the \( g \) (\( \alpha \)) and \( g-1 \) (\( \text{split} \)) parameters are those values which maximizes the likelihood function:

\[ L(\alpha_1, \alpha_2, ..., \alpha_g, \text{split}_1, \text{split}_2, ..., \text{split}_{g-1}) = L_1(\alpha_1, \text{split}_1) \cdot L_2(\alpha_2, \text{split}_2) \cdot ... \cdot L_g(\alpha_g, \text{split}_g) \]
\[ \text{where all the split parameters are non-negative (split}_m \geq 0 \text{ for } m = 1, 2, ..., g \) and sum to one (split}_1 + \text{split}_2 + ... + \text{split}_g = 1). The likelihood functions \( L_1, ..., L_g \) for the separate groups are given by Equation 5.
The corresponding gambin model of the observed frequencies is calculated as the sum of the abundances in the intervals of each group:

\[
\hat{y}_k = \sum_{m=1}^{\hat{\alpha}} \text{split}_{m, ML} \times y_{obs} \times P_y(\hat{\alpha}_{m, ML}) \quad \text{for } k = 1, 2, \ldots, \text{imax}
\]  

(7)

Whilst it is possible to use these equations to fit gambin distributions with any number of modes, in practice fitting SAD models with more than three (possibly even two depending on sample size) modes will likely result in overfitting the data; sample sizes in ecological studies are generally relatively small, and the number of parameters becomes large with increasing number of modes, and thus optimising the likelihood functions becomes increasingly problematic at larger \( g \); ecological interpretation of model fits with large numbers of modes is also problematic. As such, we do not advise fitting gambin models with more than three components.

**Using the gambin R package to fit multimodal gambin models**

Functions to fit and evaluate gambin models with \( g \) modes using maximum likelihood have been integrated into the existing gambin R package (version 2.0). Exemplar code (see also the gambin R package documentation and associated vignette) to fit and evaluate multimodal gambin models is provided below:

```r
n = 500 #set the number of species
abundances = rgambin(n, 4, 10) #take 500 random values (octave numbers/species) from a #gambin distribution with \( \alpha = 4 \) and the max octave set to 10.
abundances = rgambin(n, c(1, 4), maxoctave = c(6, 10), w = c(0.5, 0.5)) ##or take 500 values #from a two-component gambin distribution
data(moths) #load a sample dataset
fits1 = fit_abundances(data, subsample = z, no_of_components = y, cores = d)
# fit a gambin distribution with y modes to a vector of abundances, with the option of #subsampling z individuals. If y is set to 1, the standard unimodal gambin distribution is #fitted, y=2 fits the bimodal gambin distribution, and so on. To speed up computation, when #the no_of_components is greater than 1, the ‘cores’ argument can be used to enable parallel #processing using d cores.
barplot(fits1) #plot the empirical SAD
points(fits1) #add the fitted values to the barplot
AIC(fits1) #calculate the AIC value of the fitted model
```

It should be noted that it is not straightforward to directly calculate confidence intervals for the alpha parameters in gambin models with more than one mode; however, it is relatively straightforward to generate confidence intervals via bootstrapping (Puth et al., 2015).
Simulations

To test the error rates of our new models, and any sensitivity to the number of species and individuals (Locey & White, 2013), we ran three separate simulations.

**Sensitivity of model fits to the number of species**

To test how the performance of the bimodal gambin model varied with the number of species in the sample, we first sampled $n$ species from a community bimodal gambin distribution with set community parameter values $Pt$ ($alpha_1$ and $alpha_2$), using the ‘rgambin’ function. The bimodal model was fitted to this sample and the sample parameter values $Pi$ (i.e. the two alpha values) stored. This procedure was then repeated with a different value of $n$, with $n$ increasing from 20 to 1000 species in increments of 20 (i.e. up to $i = 50$). As sampling random values from a gambin distribution with set parameter values is stochastic, this incremental sampling and model fitting process was then repeated 100 times with the same $Pt$ values. We took the median sample parameter estimates of the 100 iterations (i.e. the median of the 100 parameter estimates for a given $n$) and compared these to the community parameter values ($Pt$) of the underlying distribution. It was necessary to use the median rather than the mean as when $n$ was very low the model fit would occasionally not converge and unreasonably high alpha values would be generated. For practical applications, this issue of non-convergence may be avoided by simply fixing the lower and upper bounds of the alpha parameters; for example, between 0.1 and 20. The $Pt$ alpha values were set to 1 and 4, the max octave of the simulated community was set to 13, and the weights were set to 0.5. We experimented with different community $Pt$ alpha values, but the results were qualitatively similar and so we only present the results using $Pt$ alpha values of 1 and 4, below.

**Sensitivity of model fits to the total number of individuals**

In order to test the sensitivity of the bimodal gambin model to the total number of individuals, we followed the approach of McGill (2011) and used two well specified SAD datasets from the literature: (1) the 2005 BCI tree dataset (number of species = 229; number of individuals = 20,852), comprising all trees with DBH greater than 10 cm from a 50 ha forest plot in Panama (Hubbell et al., 2005), and (2) a coral reef dataset (number of species = 154; number of individuals = 44,255) from Australia (Dornelas & Connolly, 2008). First, for each community SAD we fitted and examined the bimodal gambin model to determine whether it provided a good fit to both datasets. We then employed an iterative subsampling procedure whereby we first subsampled $x_1$ individuals from each community SAD, fitted the bimodal model and recorded the two alpha parameter ($alpha_1$ and $alpha_2$) estimates. We then increased $x_1$ by an increment $y$ and subsampled $x_2$ individuals from each community SAD, and so on until $x_i$ was equal to the number of individuals in the community SAD. We set $x_1$ to 100 for both datasets, and $y$ to 400 for the BCI dataset and 800 for the coral reef dataset, given their differences in terms of number of individuals. We then plotted the $alpha_1$ and $alpha_2$ values from these subsamples against $x_i$ to determine the influence of the number of sampled individuals on the parameter estimates. As sampling from a distribution is a random process, we re-ran the subsampling 100 times for each value of $x_i$ and took the median of the
two alpha parameter estimates (i.e. the median of 100 alpha1 values and 100 alpha2 values for each xi). For the purposes of this analysis, for the coral reef dataset we edited the maximum likelihood optimisation procedure so that both of the two compound distributions in the bimodal model had the same max octave (i.e. the max octave of the empirical distribution). This was because the maximum likelihood value without this change corresponded to a model in which one of the compound distributions only covered the first two octaves; and thus did not provide a useful example for our tests. Fixing the max octave for both compound distributions only resulted in a model with a slightly lower likelihood, but provided a better test case for the purposes of this simulation.

Determining the error rate

When working with multimodal SAD models, and ecological models more generally, it is useful to know the error rate of the model fits, i.e. the proportion of times that a multimodal model provides a better fit than a unimodal model to a unimodal dataset (i.e. false positive) and the proportion of times a unimodal model provides a better fit than a multimodal model to a multimodal dataset (i.e. false negative) (see Antão et al., 2017), and under what conditions does the error rate vary. To this end, we simulated both a unimodal and a bimodal gambin distribution with given alpha parameters and number of species (i.e. it is known a priori how many modes each distribution has). We then fitted both the unimodal and bimodal gambin models to each distribution and calculated the AIC values of the model fits. It was then possible to determine the number of times a bimodal gambin model provided a better fit than a unimodal model, based on AIC, to a unimodal sample (false negative), and vice versa. This process was repeated 100 times and the total error rate expressed as a percentage. For the unimodal simulation, the number of species, the alpha value and the max octave of the simulated datasets were set to 100, 1 and 10, respectively. The weights for the simulated bimodal distribution were set to 0.3 and 0.7 (for the first and second compound distributions, respectively), and the max octaves were set to 7 and 10 (again for the first and second distributions, respectively). We experimented with different starting alpha values and number of species to determine whether/how the error rate changes when the modes in the bimodal distribution become closer together. To achieve this, in each simulation the alpha1 value (i.e. the distribution corresponding to the relatively rarer species) was set to 0.5, whilst the alpha 2 value was iteratively changed across simulations, from 2 to 10, in units of 1. The number of species were set to 50, 100, 200 and 500.

Empirical Analyses:

In addition to the simulations described above, we also tested the multimodal gambin approach using an empirical dataset put together by one of the authors (RFK), comprising abundance records of horse flies (Diptera, Tabanidae) from a variety of locations in South America, Central America and Mexico, with a particular focus on Brazil. Data were sourced from the literature (e.g. Barbosa et al., 2005) and the subset of the dataset focused on Brazil comprised data from 33 localities across the country. Data collection is still on-going and the full dataset will be published in a subsequent paper; however, the dataset as is provides a useful test case for multimodal gambin models. Where the same species was recorded in
multiple localities, we summed the abundance values across these records for use in constructing the SAD (number of unique species = 164; total number of individuals = 78, 755). One potential reason for the prevalence of observed multimodal SADs in nature arises from the amalgamation of different samples or different types of species (Magurran & Henderson, 2003; Matthews et al., 2014b; Antão et al., 2017) within a dataset. To test this proposition, we first fitted the unimodal, bimodal and trimodal versions of gambin to the whole Brazilian dataset. We then took a subset of the dataset relating to one individual locality and one type of sampling (Centro de Instrução de Guerra na Selva, close to the Manaus city, 02°45'33''S, 59°51'03''W; canopy sampling only; number of unique species = 58; total number of individuals = 1943) and again fitted the three models. In both cases the three models were compared using AIC. All analyses were undertaken in R (version 3.4.0; R Core Team, 2017).

RESULTS

Sensitivity of model fits to the number of species

When the simulation was run using the bimodal gambin model and community alpha parameters of 1 and 4, the estimation of the alpha1 parameter (i.e. the alpha parameter of the first distribution) was relatively insensitive to the number of species in the sample (Fig. 1). That is, the sampled data provided accurate estimations of \( Pt(\alpha_1) \) regardless of the number of species. The estimation of the alpha2 parameter (i.e. the alpha parameter of the second distribution) was more sensitive when the number of species in the sample was low (Fig. 1). In particular, when the number of species in the sample was below 100, the model tended to overestimate the value of \( Pt(\alpha_2) \). However, it was only when the number of species in the sample was very low (i.e. 20) that the model substantially overestimated \( Pt(\alpha_2) \). Using different community \( Pt \) values resulted in the same pattern emerging.

Sensitivity of model fits to the number of individuals in the sample

The two alpha parameter estimates from the bimodal model were 0.50 and 6.41, and 1.07 and 11.08, for the BCI and coral datasets, respectively. Plotting the fitted values indicated that the bimodal model provided relatively good fits to both datasets (see Fig. 2a, b). Results from the subsampling analysis were similar for both the BCI dataset and the coral dataset (Fig. 2c, d). In regards to the alpha parameter of the first distribution in the bimodal model (i.e. alpha1), the parameter estimate was relatively accurate regardless of the sample size; however, at very low sample sizes the model tended to overestimate the alpha1 parameter when the BCI dataset was used (Fig. 2c). In regards to the alpha2 parameter, in both cases the parameter estimate was strongly influenced by the number of individuals in the sample, such that at low sample sizes the model estimated alpha2 parameter value varied quite considerably from the true value (Fig. 2c, d). As sample size increased, the alpha2 parameter estimate appeared to converge on the values for the full dataset; however, this convergence was not asymptotic.
How often does the model comparison analysis incorrectly determine the number of modes in the SAD?

When a unimodal gambin distribution was simulated, the error rate was 13.0%. That is, on average, the unimodal model provided a better fit than the bimodal model 87.0% of the time. Changing the number of species in the simulation did not affect the qualitative results (results not shown).

When a bimodal gambin distribution was simulated (number of species varied between 50 and 500; alpha1 kept constant at 0.5; alpha2 varied from 2 to 10), the mean error rate depended on the sample size and the difference between the alpha1 and alpha2 values (Fig. 3). When the difference between alpha1 and alpha2 was relatively large, the error rate (i.e. the proportion of times AIC selected the bimodal gambin model as the best model) was very low (e.g. 0%) regardless of the number of species in the sample. In contrast, when the difference between the alpha1 and alpha2 values was very small, the error rate was high (e.g. 80%) regardless of the number of species. In between these extremes, the error rate at a given level of difference between alpha1 and alpha2 depended to some extent on the number of species (Fig. 3).

Fitting multimodal gambin models to a South American horse fly dataset

When the three models were fitted to the whole Brazilian horse fly dataset (Fig. 4a), the bimodal gambin model provided the best fit to the data (AIC = 827.2), followed by the unimodal model (AIC = 831.2) and the trimodal model (AIC = 832.5). When the three models were fitted to the subset of data from Centro de Instrução de Guerra na Selva (Fig. 4b), the unimodal model provided the best fit (AIC = 235.8), followed by the bimodal model (AIC = 238.9) and the trimodal model (AIC = 244.9).

DISCUSSION

In this paper, we have derived the maximum likelihood equations for gambin models with multiple modes and integrated these functions into an updated version of the ‘gambin’ R package which has been uploaded to CRAN (the original version is outlined in Matthews et al., 2014a). We have shown that, in general, the new multimodal gambin models perform adequately under a variety of circumstances, and that the application of our new models to empirical SAD datasets can provide interesting insights. However, there are a number of cases (e.g. low numbers of individuals in a sample), in which the model performance is questionable and we discuss these below.

The results of our simulations involving varying the number of species in a sample indicated that in general the alpha parameter estimates of the bimodal gambin model were relatively insensitive to the number of species in the sample. There was some bias (see also McGill, 2011 for a discussion of this issue in SAD analyses more generally), such that at low sample species richness the alpha2 parameter tended to be overestimated. However, it was only when species richness was very low (< 50) that the sample alpha2 value diverged substantially from the true alpha2 value. The reason for this observation is likely due to the fact that
reducing richness in the simulated samples results in a concomitant reduction in the number of individuals in the sample (discussed further below).

In contrast to the number of species in the sample, it is apparent that to get truly accurate parameter estimates for the bimodal model fits, the number of individuals in the sample needs to represent a large proportion of the number of individuals in the community. This is true of most SAD models and is worrying given that this criterion is rarely met in SAD analyses. For example, McGill (2011, p.*) found, using a similar subsampling method, that “most metrics only began to come close to their true values (even to within +- 50%) with at least 1000 individuals sampled and in many cases only with 10,000 individuals sampled.” The alpha parameter from the unimodal gambin model was determined to require sample sizes of 1000 individuals (McGill, 2011). Furthermore, McGill’s analysis was focused primarily on unimodal distributions, and arguably the problem can be seen to be even more acute for compound distributions, due to the generally larger number of parameters and more complex optimisation algorithms needed to find the maximum likelihood estimates of the parameters.

In regards to our simulations using the bimodal gambin model, it is the alpha2 parameter that is most sensitive to sample size, and although alpha2 increases with increasing sample size, albeit with scatter at very small sample sizes, it does not perfectly asymptotically converge on the true value. This issue appears to be largely caused by sampling problems. Primarily, when smaller and smaller samples are taken from a community SAD that is distinctly multimodal, the multiple modes become more difficult to detect. Thus, this is a problem for all multimodal SAD models, although it is potentially more of an issue with gambin due to the prior binning of the data. Figure 5 illustrates this sampling effect. In Fig.5d the fit of the bimodal gambin model to the BCI dataset is shown. The fit of the bimodal model to samples of 100, 1000 and 10,000 individuals from the BCI dataset (Fig. 5a, b, c) are also shown. It can be seen that, as sample size decreases, the second mode (i.e. the mode representing the relatively more common species) shifts to the left, whilst the first mode (i.e. the mode representing the singleton class) remains static (Fig. 5). This fits with Preston’s (1948) concept of the veil line; it is only by increasing the size of the sample that the full empirical distribution is revealed. This explains why the alpha1 parameter is relatively insensitive to variations in sample size, whilst the shifting of the second mode to the left of the distribution explains why alpha2 tends to decrease with sample size. The shifting of the second mode with variation in sample size is not a ‘smooth’ continuous process, which is why alpha2 does not converge towards the true value in a smooth asymptotic fashion. The reason for this relates to the changing empirical max octave value of the sample. Starting with a small sample size (i.e. a low proportion of the number of individuals in the community), as the number of sampled individuals increases at some point a new max octave is added (i.e. there are species now in the sample with higher abundance than in samples of smaller sizes; e.g. compare the different empirical max octaves in Fig. 5a-d). The addition of a new max octave acts to stretch out the gambin distribution which in turn influences the alpha values.

The sensitivity of the alpha parameter to sample size was found in previous work focused on the unimodal gambin model (Matthews et al., 2014a). However, in contrast to the unimodal model, in regards to the bimodal model we do not think that this sensitivity is necessarily a
substantive issue, depending on the aim of the study. With the unimodal gambin model, the
alpha value can be used as a type of diversity metric to compare SAD shape across
communities (e.g. Arellano et al., 2017). Thus, we have previously advised that subsampling
should be used to ensure constant sample size in comparative analyses using the unimodal
model (Matthews et al., 2014a). However, for multimodal gambin models the meaning of the
alpha values is not as clear; partly due to the fact that the max octave of the first distribution
(in the context of the bimodal model) is allowed to vary and thus will vary between different
samples. As such, when fitting multimodal gambin models we do not advise using the alpha
parameter estimates as diversity metrics or as response variables in regression-type
comparative analyses. Rather, the benefit of multimodal gambin models is to provide a
simple, quick and easy to use test for determining whether empirical SADs are multimodal;
the alpha parameter values can then be used to simply provide a rough idea of the shape of
the compound distributions in certain cases.

Furthermore, according to our simulations, unless sample size is very low (Fig. 2) alpha2 is
still relatively large compared to alpha1 and thus indicates to the user the presence of two
modes in the distribution. It is also worth noting that, as gambin is a statistical distribution
rather than an ecological model, obtaining exact parameter estimates, particularly for
multimodal models, is arguably less important than for other, ecological models (e.g. the
migration parameter, m, in various neutral models; e.g. Hubbell, 2001; Chust et al., 2013).

In regards to our error rate simulations and the bimodal SAD example (i.e. when the
simulated distribution was bimodal), the fact that the error rate increases as the modes
become closer together (Fig. 3) is to be expected, as the underlying sample distribution starts
to resemble a unimodal distribution. For example, when the two alpha values are 0.5 and 1
the resultant simulated bimodal gambin distribution is visually unimodal and thus the high
error rate here is not surprising. As most empirical multimodal SADs have distinct rarer and
more common species modes, this is not a substantive issue. In addition, the fact that the
model comparison test is biased towards selecting the unimodal model over the multimodal
model and not vice versa is favourable as it indicates that using gambin models to test for
SAD multimodality (in combination with information theoretic model comparison
approaches; Burnham & Anderson, 2002) is a conservative approach. That is, the error rate
for erroneously classifying a SAD as being multimodal, which is generally the alternative
hypothesis in multimodal SAD studies (e.g. Matthews et al., 2014b), is low.

Empirical analyses
The analyses using the horse fly dataset indicated that when the different Brazilian records
(i.e. records from different localities across Brazil) were pooled, the resultant SAD was
clearly multimodal (see Figure 4a). This is likely to be due to the fact that this dataset
comprises data from multiple studies and between-study differences in sampling
methodology, location and study aim mean that the pooled dataset may in fact comprise
multiple individual SADs. In particular, differences in sampling methodology (e.g. the type
of traps used) may result in different types of species being sampled between studies. The
amalgamation of different types of species (i.e. increasing taxonomic breadth) has been
shown to affect the shape of the SAD (e.g. Magurran & Henderson, 2003), and multimodal SADs have been found to be more prevalent with increasing taxonomic breadth and spatial extent (i.e. increasing ecological heterogeneity; Antão et al., 2017). Thus, the amalgamation of different samples is likely the cause of the multimodal SAD we observed for the Brazilian dataset. In contrast, when the subset data (i.e. data from one locality within Brazil using only samples from the canopy) were used, evidence of multimodality was less apparent (Fig. 4b). However, the SAD is clearly not perfectly unimodal either (Fig. 4b). The reasons for this finding are unclear. Other mechanisms have been argued to produce multimodal SADs, such as emergent neutrality (Vergnon et al., 2012; see Matthews & Whittaker, 2015) and differences in dispersal ability (Borda-de-Água et al., 2017); however, testing these different hypotheses is beyond the scope of this paper. Regardless, these analyses highlight how fitting and plotting the fitted values of the unimodal and multimodal gambin models provides a quick and easy method for analysing these types of questions.

Conclusions and future opportunities

Multimodal gambin models represent a novel, quick and easy test for determining whether SADs are multimodal. Due to the relatively simple underlying mathematics and binning procedure, and the relatively small number of parameters compared to other compound distributions, the models are easy to fit and the maximum likelihood estimation procedure does not require the user to vary the staring parameter values or the optimisation algorithm employed. Furthermore, the alpha parameters can provide a rough summary of the shape of the SAD. Previous ecological interpretations of alpha, such as niche dimensionality (Ugland et al., 2007), are less clear in the context of multimodal gambin models and require further research.

Whilst our method of determining multimodality involved the use of an information theoretic model comparison criterion (i.e. AIC), it would be possible to use a variety of different methodological approaches with this aim in mind. For example, as gambin models with increasing numbers of modes are nested, it would be possible to use likelihood ratio tests in addition to information theoretic criteria. Other more stringent criteria could involve using multiple criteria simultaneously, in addition to approaches such as parametric bootstrapping (Connolly & Dornelas, 2011; Matthews & Whittaker, 2014; Antão et al., 2017). In addition, whilst gambin models have so far only been used to analyse SADs (e.g. Ugland et al., 2007), it is possible to fit them to any other type of ecological or general distribution. In regards to multimodal gambin models, it would be interesting to assess their performance and alpha estimates when fitted to species-range size distributions and abundance/species-occupancy distributions (Hanski, 1982; Gaston & Blackburn, 2000; Borregaard & Rahbek, 2010). For example, there is evidence that some species-range size distributions may exhibit multimodality (e.g. see Gaston, 2003, p. 80).

As Antão et al. (2017, p. 203) state, “multimodality occurs with a prevalence that warrants its systematic consideration when assessing SAD shape and emphasizes the need for macroecological theories to include multimodality in the range of SADs they predict.” The development of multimodal gambin models provides one tool to undertake these types of
analyses. Application of these new models to additional datasets will likely be revealing and will help in improving our understanding of multimodal SADs, and SADs more generally.

ACKNOWLEDGEMENTS

The BCI forest dynamics research project was made possible by National Science Foundation grants to Stephen P. Hubbell: DEB-0640386, DEB-0425651, DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042, DEB-8206992, DEB-7922197, support from the Center for Tropical Forest Science, the Smithsonian Tropical Research Institute, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Small World Institute Fund, and numerous private individuals, and through the hard work of over 100 people from 10 countries over the past two decades. The plot project is part of the Center for Tropical Forest Science, a global network of large-scale demographic tree plots.

REFERENCES


Figure 1 Variation in the two alpha parameter estimates (alpha1 = black circles; alpha2 = red circles) of a bimodal gambin model with the number of species in the sample. Samples were taken from a bimodal distribution with alpha1 and alpha2 parameters set to 1 and 4 (solid black lines), respectively. For each number of species value, the subsampling was repeated 100 times and the median alpha1 and alpha2 values taken.
Figure 2 The fits of the unimodal (blue circles) and bimodal (red triangles) gambin models to two empirical SADs: (a) the 2005 BCI tree dataset (number of species = 229; number of individuals = 20,852) from Panama, and (b) a coral reef dataset (number of species = 154; number of individuals = 44,255) from Australia. Subsamples of these datasets were created by subsampling a varying number of individuals (NI) in each case. The bimodal gambin model was fitted to each subsample and the average alpha1 (black circles) and alpha2 (red circles) parameters (average of 100 iterations for each NI value) were stored. These values are plotted against NI for the BCI data (c) and the coral data (d). The blue diamonds in (c) and (d) are the true parameter values, i.e. the parameters of the model fits to (a) and (b). The bimodal model in (b) was fitted using the standard function in the gambin R package that allows the max octave of the first component distribution to vary (i.e. it does not have to equal to max octave of the empirical distribution). However, to enable a better test of the effect of the NI on the alpha estimates (d), it was necessary to use a different function that
fixed the max octave of both component distributions (see ‘Methods’ for further information).
Figure 3 The multimodal SAD error rate (expressed as a percentage) for an information theoretic model comparison test. For the test, a bimodal SAD was simulated, with one alpha parameter fixed at 0.5 and the second set to vary between 2 and 10 in units of 1. The number of species (sample size) was set to: 50, 100, 200, 500. The unimodal and bimodal gambin models were then fitted to this simulated SAD and the best model fit determined using AIC. The error rate percentage relates to the proportion of times the unimodal model provided a better fit than the bimodal model (i.e. a higher error rate percentage indicates that the unimodal model erroneously provided a better fit to the bimodal SAD).
Figure 4 The fit of the unimodal (blue circles), bimodal (red triangles) and trimodal (black diamonds) gambin models to two horse fly species abundance distribution datasets (black bars) from South America. (a) horse fly data from 33 localities across Brazil (number of unique species = 164; total number of individuals = 78,755), and (b) data from one individual locality and one type of sampling (Centro de Instrução de Guerra na Selva, close to the Manaus city; canopy sampling only; number of unique species = 58; total number of individuals = 1943). In (a) the bimodal model provides the best fit according to AIC, whilst the unimodal model provided the best to (b).
Figure 5 The fit of the bimodal (red triangles) gambin model to the SAD (d) of the 2005 BCI tree dataset (number of individuals = 20,852), and to three samples from this dataset of varying size: (a) 100 individuals, (b) 1000 individuals, and (c) 10,000 individuals.